1-1-2005

Cross layer design and performance analysis of 3G cellular CDMA downlinks

Jin Yuan Sun
Ryerson University

CROSS LAYER DESIGN AND PERFORMANCE ANALYSIS OF 3G CELLULAR CDMA DOWNLINKS

by

Jin Yuan Sun
Bachelor of Science Degree in Computer Information Systems,
China, 2003

A thesis
presented to Ryerson University
in partial fulfillment of the
requirements for the degree of
Master of Applied Science
in the Program of
Computer Networks

Toronto, Ontario, Canada, 2005

Ryerson University requires the signatures of all persons using or photocopying this thesis. Please sign below, and give address and date.

Cross Layer Design and Performance Analysis of 3G Cellular CDMA Downlinks, Master of Applied Science, 2005, Jin Yuan Sun, Computer Networks, Ryerson University

Abstract

In this thesis, a cellular CDMA network with voice and data communications is considered. Focusing on the downlink direction, we seek for the overall performance improvement which can be achieved by cross layer analysis and design, taking physical layer, link layer, network layer and transport layer into account. We are concerned with the role of each single layer as well as the interaction among layers, and propose algorithms/schemes accordingly to improve the system performance. These proposals include adaptive scheduling for link layer, priority-based handoff strategy for network admission control, and proposals for the avoidance of TCP spurious timeouts at the transport layer. Numerical results show the performance gain of each proposed scheme over independent performance of an individual layer in the wireless mobile network. We conclude that the system performance in terms of capacity, throughput, dropping probability, outage, power efficiency, delay and fairness, can be enhanced by jointly considering the interactions across layers.

Acknowledgments

The author wishes to send her appreciations to Dr. Zhao whose supervision and support greatly assist on the fulfillment of this thesis. The author would like to thank other committee reviewers Dr. Anpalagan, Dr. Jaseemuddin, and Dr. Ma for their valuable comments to improve the quality of this thesis. Thanks are also due to Mr. Yongyu Jia for his proofreading and my family members Donglin, Jianxin for their endless support.

Contents

List of Tables

List of Figures

Chapter 1

Background And Objectives

Code Division Multiple Access (CDMA) is a technology used in the air interface and is categorized as a spread spectrum technology. The fundamental philosophy of the spread spectrum technology is that the information symbols of a narrow bandwidth are multiplied by high chip rate spreading codes (pseudo-random codes) and result in a spread signal with wide bandwidth. The capacity of the spread bandwidth relies on the number of chips per symbol. A remarkable advantage of this technology is the enhancement of the robustness of the spread signal to interference and distortion, compared to the original narrowband signal. Earlier wireless Medium Access Control (MAC) technologies such as Time Division Multiple Access (TDMA) and Frequency Division Multiple Access (FDMA) have a common limitation of resource availability (*i.e.*, time-slots in TDMA and frequencies in FDMA), imposing contention among users, restricting network capacity and resource utilization since no more than one user can access the medium at the same time (TDMA) or occupying the same frequency (FDMA). CDMA has attracted attention because it overcomes the above problem by assigning each transmitting user a unique "code channel", with which users can access the medium simultaneously using the entire bandwidth because they will appear as noise to one another without collision. Moreover, the spread-spectrum basis of CDMA technology further increases the system capacity due to the interference-resistant nature. We propose to use CDMA systems as the basis of our research. The simulations are conducted in such an environment and performances are analyzed accordingly.

1.1 Background

The hierarchical structure of the wired networks, which is the well-known seven-layer OSI (Open System Interconnection) reference model, consists of physical layer, data link layer, network layer, transport layer, session layer, presentation layer and application layer, from bottom to top. The rationale behind this design is for easier implementation of peer-to-peer communication since each layer has clear responsibility and hides internal functions from others. Fig. 1.1 shows the communication between two hosts A, and B, across the seven layers of each host.

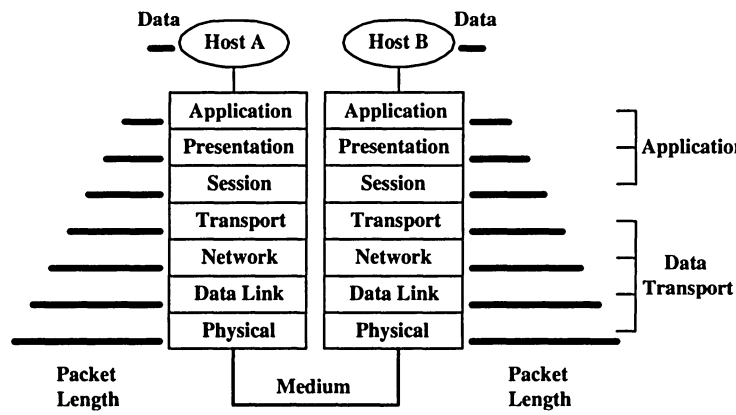

Figure 1.1: The OSI Reference Model.

Note that this layered model is just a conceptual framework and is not for actual communication itself. Instead, each layer employs certain protocols which are in charge of the peer-to-peer communication and inter-layer interaction. Suppose that Host A has information to send to Host B. The application layer of A passes the data all the way down to the physical layer of A. During this procedure, each lower layer attaches its control information (header/trailer) to the data passed down from the upper layer, which is known as encapsulation. The physical layer of A then transmits the data through physical medium to B. The control information attached will be used by each peer layer of B to decapsulate the data packet as it is passed from the physical layer to the application layer of B, until finally received by B with the original data. We briefly introduce the functionality of

these layers. The physical layer defines the electrical, mechanical, procedural, and functional specifications (*i.e.*, voltage levels, timing of voltage changes, physical data rates, maximum transmission distances, and physical connectors) for activating, maintaining, and deactivating the physical link between communicating network systems; the data link layer defines data format and provides reliable transit of data across a physical network link; the network layer defines the network address and the logical network layout, where routers determine the way to forward packets; the transport layer accepts data from the session layer and segments the data for transport across the network. It ensures error-free and in-order data delivery, and provides flow control and congestion avoidance; the session layer establishes, manages, and terminates communication sessions; the presentation layer provides a variety of coding and conversion functions that are applied to application layer data; the application layer interacts with software applications that implement a communicating component. The upper three layers handle application issues while the lower four layers deal with data transport, which is the highlight of our study.

The conventional structure of wired network also divides wireless network into layers, hence the wireless network has the same layered structure as the wired network. However, the tasks performed by each layer in wireless network may vary. We explain the duties of the four layers involved in data transport in a CDMA network. The physical layer deals with the wireless channel, which features time-varying characteristics and impairments (path loss, shadowing, and fast fading) incurring high bit error rate and under-utilization of the system capacity. The task of the physical layer is to combat these adverse effects on data transmissions and to increase resource utility. Power control technique is therefore applied to compensate channel impairments in CDMA downlinks (while uplinks use power control mainly for eliminating the near-far problem), while power/rate allocation (distribution) is proposed for capacity and throughput maximization. The link layer relies on the packet scheduling technique to dynamically adjust the transmission rates and coordinate packet transmissions from different users for efficient utilization of the system resources based on the physical-layer channel conditions. Admission control is the network-layer mechanism which

determines the admission of new and handoff connections for the system to accommodate larger number of users, under the restriction of certain criteria (*i.e.*, different treatments with new/handoff connections). The transport layer manipulates the end-to-end reliable data transport across wired/wireless integrated network, maintains the flow/congestion control mechanism, and renders TCP-level throughput not to be degraded in the presence of wireless impacts.

1.2 Motivations and Objectives

With the growing demand and popularity of high-speed data applications in wireless networks, the system capacity and bandwidth resource become increasingly stringent. Thus research efforts are made to expand the capacity and to efficiently allocate the resource. Some of the efforts include the evolution of the CDMA technology. For example, 3G (third-generation) and 3G+ (beyond 3G) CDMA systems (*i.e.*, WCDMA, cdma2000) are superior to 2G CDMA systems (IS-95) in that the former ones have higher carrier bandwidth and faster power control frequency (more precise channel feedback). On the other hand, some efforts focus on the algorithm design which can be implemented in the software to optimize the system performance "softly". As the research on 3G and 3G+ CDMA systems emerges, and although countless works have contributed to this research area, there still remains a great number of problems unsolved. Especially when nowadays the speedy development and upgrade of the information technology fuel the update and renewal of techniques and skills, more new challenges appear without the old ones being solved. In this thesis, we would like to share our ideas to approach some of the problems in this field. We choose to focus on the cross layer design of the CDMA networks as this issue has recently been capturing interests.

The traditional wireless networks mainly support voice service without data service provided through the Internet backbone. However, the integration of the wireless network and the wired backbone is of great importance today because of the increasing data application requirement at the mobile terminal (*e.g.*, cellular phone, wireless laptop). While most previous research was on the performance optimization of individual layer, it often leads

to performance degradation of other layers or suboptimal system performance. The hierarchical structure of the wireless networks, as the wired ones, facilitates us to design and study protocols for the single layer that is of particular interest since these layers (physical layer, link layer, network layer, transport layer and application layer) are transparent to one another. But this isolation may cause suboptimal system performance. Recent research has shown that a well-designed cross-layer approach that supports multiple protocol layer adaptivity and optimization can yield significant performance gains [1]. Many researchers use the cross-layer approach for their designs. However, these designs can be very different due to various combinations and interactions of multiple layers.

We have studied a number of cross-layer approaches for CDMA system optimization in the literature. In what follows, we summarize some of these approaches. [2]-[8] propose cross-layer approaches to achieve system optimization in CDMA systems. In [2] a set of PHY-MAC (Physical-MAC) mechanisms is proposed based on the rate adaptation provided by the MAC and the channel state from the PHY to improve spectrum efficiency and reduce power consumption. Yu and Krishnamurthy [3] focus on cross-layer QoS (quality-of-service) guarantee by combining physical layer SIR (signal-to-interference-ratio) and network layer blocking probability to reduce computational complexity and approximate the optimal solutions. Other works are also found to address physical/network cross-layer optimization issues [9]. Price and Javidi [4] deal with the interaction between congestion (transport layer) and interference (MAC layer), and integrate them into a single protocol by means of rate assignment optimization. Friderikos *et al.* [5] interpret the rate adaptation as TCP-related since the rate in this paper is defined as the ratio of the current congestion window and RTT (round-trip-time) of the connection, and jointly considers it with physical layer (power). Hossain and Bhargava [6] model and analyze the link/PHY level influence on TCP behavior and illustrates their dependency. Yao, Wong and Chew [7] study the reverse and forward link capacities balancing issue by covering link layer and the network layer to seek for optimal handoff probability. Chan *et al.* [8] propose a joint source coding power control and source channel coding, and interpret them as the MAC-layer power control and application-layer

source coding, respectively, maximizing the delivered service quality and minimizing the resource consumption. There are also additional attention on other aspects of cross-layer design, such as to decrease the cross-layer interference [10] instead of optimization.

We can see from the above that different interpretations of "cross layer" and resources belonging to these layers produce a variety of cross-layer studies. While existing works address cross-layer issues based on two or three layers, we propose to fully address this issue by taking the four important layers into account: physical layer, link layer, network layer and transport layer. We design algorithms/protocols for each of these layers by considering their communications and mutual impacts to prevent isolation thus improving the overall performance.

It should be noticed that cross-layer design is suitable for wireless mobile networks because in general it has some advantages in such networks. Compared with layered design, the cross-layer design exploits inter-layer interactions, achieves protocol optimization under system constraints, improves the adaptability and performances across layers through exchanged information. On the other hand, layered design generally applies to large and reliable networks, where it reduces the design complexity, improves the maintainability and keeps the modularity. Moreover, it is easy to standardize and flexible to deploy new protocols. These merits of the layered design do not exist in the cross-layer design. In addition, the cross-layer design may have some other drawbacks that need design cautions. First, joint optimization across layers yields more complex algorithms and will probably cause unnecessary optimization affecting the regular functionality of the layer. Second, design loops may occur with incautious design. Also, it is hard to characterize the interactions between protocols at different layers and is likely to destroy the modularity. Based on the explanations about the advantages and drawbacks above, we conclude that to successfully design a cross-layer approach, the core issue is to understand and exploit the interactions among layers. This is the main thread and foundation in all the proposed cross-layer approaches in this thesis.

1.3 Thesis Outline and Contributions

The main body consists of Chapters 2, 3, 4, and 5, where we propose strategies for physical layer, link layer, network layer, and transport layer, respectively, study their interactions, and analyze the performances of these proposals.

- Chapter 2 first illustrates the simulation environment and the system model based on which this research is carried out. It then starts introducing the first layer involved, the physical layer, in which two fundamental techniques, power control and rate allocation are studied, separately and jointly and a survey is given through a large amount of literature review. The proposed integrated power control and rate allocation is briefly introduced which is primarily used for the link-layer scheduling and is demonstrated in detail in the following scheduling schemes.

- At the link layer in Chapter 3, a novel voice packet scheduling scheme named Modified Adaptive Priority Queuing (MAPQ) and a unified framework (UF) for scheduling hybrid voice and data traffic are proposed. An adaptive priority profile is defined in these schemes based on queuing delay and physical layer information such as required transmission power, and available transmission rate, which borrows the idea of composite metric from wired systems. Estimation error is considered when measuring received pilots at mobile stations. Users are allocated resources according to their priorities in a modified PQ fashion constrained by total power budget of base stations. For MAPQ, this definition ensures system capacity improvement, packet dropping probability reduction, and fairness (in terms of the mean delay). Numerical results show distinct performance gains of the proposed scheme, comparing to the reference schemes. For UF, we address the consistency of the framework as well as the distinctions of voice and data scheduling processes by discussing the common policy and individual requirements of both classes. The uniformity of the proposed framework not only simplifies the implementation of the scheduling algorithms at base stations, but also is verified to be robust and resistant to various offered traffic load and variable service structure

(voice/data proportion). With this design, the proposed algorithm accomplishes system performance enhancement as a whole while retaining separate performance features without degradation. Numerical conclusions show a system capacity gain of 4%-39% and traffic throughput improvement of 5%-26% most of the time over first-in-first-out systems.

- Chapter 4 addresses the network-layer issues where, we propose an adaptive prioritizing soft handoff algorithm for concurrent handoff requests aiming at a same cell. A predicted set, an adaptive priority profile jointly exploiting the impact of required handoff power and call holding time have been developed to realize the proposed algorithm. A link-layer scheduler residing in each base station to ensure the desired operation of the prioritizing procedure is also designed, with input information from network layer. Numerical results are acquired through comparing the proposed algorithm with a fixed guard capacity scheme and performance gain is obtained in terms of handoff dropping probability, average guard power utilization, and average guard power efficiency, by no less than 24% and 5%, respectively for the first two criteria, and some amount for the last.

- As the upmost layer for data transport, the transport layer is concentrated on in Chapter 5. We study the problem of TCP over wireless links and summarize the solutions for this problem from extensive research works. We design and propose two alternative solutions to prevent the spurious timeouts at TCP sources caused by the stochastic intervals of wireless opportunistic scheduling. The proposals are claimed to be feasible and effective through the analysis of the rationality, the description of the implementation details, and the experiments and simulations. The results obtained for the first proposal in terms of TCP congestion window evolution and TCP throughput look convincing.

- Finally Chapter 6 concludes with open issues and certain limitations of this thesis.

Chapter 2

System Overview and Physical Layer Techniques

Wideband CDMA (WCDMA) is an air-interface technology which is the basis of UTRA (Universal Terrestrial Radio Access). UTRA is the third generation radio network technology specified by the 3rd Generation Partnership Project (3GPP), a joint standardization project consisting of standardization bodies from Europe, Japan, Korea, America, and China. The simulations of the following chapters are set up in a WCDMA environment. Before entering the proposed cross-layer design, we would like to describe some characteristics and specifications of the WCDMA standard that are related to this work.

2.1 System Overview

2.1.1 WCDMA Specifications

WCDMA technology multiplies users' information symbols by pseudo-random sequences, which have a chip rate of $3.84 Mcps$ (millions of chips per second). As a result, the narrowband information bits are spread over a wider bandwidth of up to $5.2 MHz$. Some of the key technical specifications of WCDMA used for our simulation environment setup are listed in Table 2.1.

Table 2.1: Key Technical Specifications of WCDMA

Multiple access technique	Direct-spread code division multiple access (DS-CDMA)
Frequency reuse	1
Carrier bandwidth	4.4 MHz - 5.2 MHz
Chip rate of spreading bits	3.84 Mcps
Data type	Packet and circuit switching
Maximum user data rate	2.3 Mbps
Frame length	10 ms (38400 chips)
Number of slots / frame	15
Number of chips / slot	2560
Intra-system handoff	soft / softer handoff
Power control period	Time slot = 1500 Hz rate
Power control step size	0.5, 1, 1.5, 2 dB (variable)
Physical layer spreading factors	4...256 (Uplink) 4...512 (Downlink)

2.1.2 System Model

We study a cellular WCDMA system with wrap-around cell structure, as shown in Fig. 2.1, where mobile stations (MSs) select serving base stations (BSs, displayed as pentagrams) based on the measured strength of pilot signals (P_p) sent out periodically by BSs. Typically, 20% of the downlink total power will be assigned to pilot channel by each BS. The rest of the total power is shared by traffic channel and other control channels (carries control information, power control symbols, *etc.*).

In forward link (downlink), BSs transmit packets to MSs through traffic channels which consist of frames. Each frame is of a length $10ms$ and is sub-divided by 15 time slots. Within one of each time slot traffic destined for a specific MS is delivered.

Upon receiving the pilot signals, MSs compare the received SIR to a target, SIR^*, and generate a power control command for feedback. BSs adjust the transmission power for MSs depending on the power control commands fed back through closed-loop power control mechanism.

To guarantee that MSs receive packets correctly, the transmission power allotted by BSs has to overcome channel impairments, which consists of path loss ($(d_{ib}/d_0)^{-\alpha}$), shadowing ($e^{-(\beta X_b)}$), and Rayleigh (fast) fading (simulated using Jakes' fading model [11]), where d_{ib} is

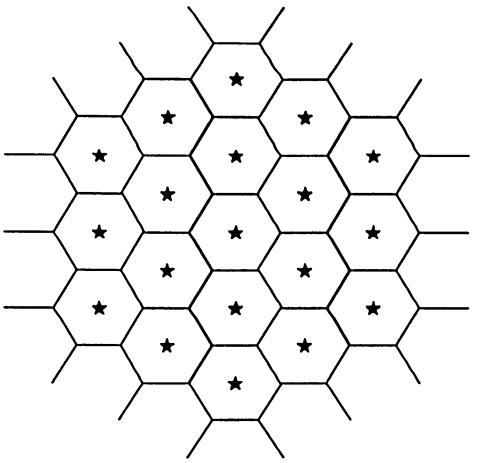

Figure 2.1: A Typical Cellular Structure with Base Stations.

the distance of user i from base station b, d_0 is the close-in reference distance, α is the path loss exponent, $\beta = ln10/10$ is a constant, and X_b is a Gaussian distributed shadowing (in cell b) random variable with zero mean and variance σ_X^2, assuming $\sigma_{X_b} = \sigma_X$, $\forall_b$.

Unlike the uplink case, downlink restraint is not the intra-cell (from adjacent MSs within home cell) and inter-cell (from neighboring other cells) interference, but the total transmission power of BSs. The E_b/I_0 (bit energy to interference density ratio), received at each mobile i from its home BS b is expressed as:

$$\gamma_i = \Gamma \frac{P_i G_{bi}}{(P_T - P_i)G_{bi} + \sum_{j \neq b} P_T G_{ji}}, \tag{2.1}$$

where $\Gamma = W/R$ is the spreading gain with W, the spread spectrum bandwidth and R, the required packet transmission rate. P_i and P_T (in $Watt$) denote the transmission power for user i and total downlink transmission power of BSs, respectively, assuming BSs are transmitting at the maximum capacity always [12]. G_{ji} denotes signal attenuation of the channel from BS j to user i. During each time slot, MS i compares the received γ_i with a target E_b/I_0, γ^*, and generates the power control command informing its serving BS b to increase or decrease the transmission power by a fixed step size.

In reality, however, limited measuring ability of the MS can introduce an erroneous estimation of received pilots, which results in errors in calculating G_{ji} (obtained from measuring

pilot signals). We approximate the sum of measurement errors as log-normally distributed as indicated in [12]. Let $e^{\beta Y_{ib}}$ denote the measurement error of the measured value, where Y_{ib} is a Gaussian distributed random variable with zero mean and a variance of σ_Y^2 (same assumption holds as to σ_X^2). As a result, the actual received SIR for each MS is derived from the following revised version of (3.1) counting errors:

$$\gamma_i = \Gamma \frac{P_i/P_T}{(1 - P_i/P_T) + e^{-\beta Y_{ib}} \sum_{j \neq b} G_{ji}/G_{bi}}. \tag{2.2}$$

2.2 Physical Layer: Power Control and Rate Allocation Techniques

2.2.1 Background

As a high-flexible and multipath-resistant technique, CDMA has attracted interests for wireless multimedia applications [13]. Second-generation CDMA systems such as those based on the IS-95 standard, are designed mainly for real-time, constant-bit-rate voice communications. In general, voice transmission has low data rate requirements with real-time delay constraints. This suggests that only power control with fixed-rate transmission is well suited to voice. Thus for such systems, transmission power control is employed to combat channel fading of the wireless link and to maintain the received power of the mobile station at a desired level. Power control is an essential technique for CDMA systems to increase the capacity. The reverse link of CDMA cellular systems suffers from the near-far effect which exists when all users send at a fixed power level. Then the signal received from a nearby user will be stronger and thus mask the signal received from a far user. Power control is used to adjust the power levels sent by different users to achieve the same received power level at the base station.

Next-generation wireless communication systems (WCDMA, cdma2000)are expected to support multimedia services, such as voice, data, video, *etc.* These services are characterized by different QoS requirements such as minimum transmission rates. While 2G systems are mainly designed to support voice applications, 3G systems are to provide high speed

data services. Data traffic has burst in nature and exhibits variable bit rates from time to time. In general, non-real time delay insensitive data transmission demands higher rates with less stringent delay requirements. This suggests that variable rate transmission, which saves transmission power and also produces less interference to other users, is best suited to data applications. For instance, email and file transfer data services may temporarily lower their transmission rates even to zero under bad channel condition and utilize any excess capacity in a best effort fashion [14] to maximize the system throughput. In order to utilize the available variable transmission rates and control them in a spectrally efficient way, a new technique named rate control (rate allocation) has emerged. Since power control is fundamental to wireless CDMA systems for voice communications and rate allocation contributes to throughput enhancement for data communications, it is natural to integrate power control and rate allocation to provide QoS and further improve the system capacity in the 3G integrated voice and data transmission systems.

Power control and rate allocation can be performed in both uplink and downlink directions. It is of great significance for the uplink because of the near-far effect mentioned above. On the other hand, it also can be very useful in the downlink because it provides an efficient way to track the time-varying channel condition and save the total power budget of the base station by transmitting at the minimum required power levels at all times, resulting in an increase in downlink system capacity. Power control is originally designed for the uplink purpose and is thoroughly studied. In this chapter, we first present a survey on power control and rate allocation techniques through an extensive review of the literature. The survey is by no means exhaustive. We seek for the basic understanding and a general portrayal of this crucial mechanism. The necessity of power control and rate allocation for downlinks in our work is also addressed.

2.2.2 Power Control Techniques

Transmission power control is a central technique for resource allocation and interference management in wireless CDMA systems. Well-defined power control is essential for proper

functioning of CDMA systems [15]. Without power control, the near-far problem is dominant and the system capacity is very low. When power control is employed, it is possible for mobile users to share the system resources equally. The objective of power control is mainly two-fold. It assigns transmission power within the power budget to each user so that the QoS requirements (*i.e.*, SIR) of all the transmissions are met. It also minimizes the total transmitting power of users to prolong the battery life [16].

Power control techniques for wireless CDMA networks have received much attention in recent years. There have been numerous proposals for power control reported in the literature as discussed below. To our knowledge, power control techniques can be classified in many different ways. There is centralized and distributed power control. The centralized power control requires the controller to get all the information of all users, such as signal strength, channel gains to control the power levels for different users. It is extremely difficult to apply centralized power control in practical systems in that this mechanism requires extensive control signaling in the network. An example of this method is given in [17]. The distributed algorithms require only the local information, such as SIR, channel gain of a specific user. Such distributed algorithms can be found in [18] and [19]. However, there are some limits in applying the algorithm in a practical system. One possible limit is the measuring and control signaling time, which results in feedback delays in the system.

Power control techniques can also be classified based on different QoS measures at the base station. Such techniques include strength-based, SIR-based and BER-based power control. SIR-based power control is extensively used in the research [17]. A serious problem associated with SIR-based power control is the potential to increase transmission power to endanger the stability of the system [15], which is the so-called non-cooperative N-person game problem [20]. Authors of [21] suggest a combination of strength-based and SIR-based schemes. Both received power and SIR are measured at the base station. This combined technique is proven to yield better performance than SIR-based technique only.

Yet, another possible way to classify power control techniques is based on whether they are open-loop or closed-loop power control. The base station transmits a pilot signal on the

forward link to assist the mobile to estimate the path gain by measuring the strength of the pilot signal in an open-loop power control [22]. This is used to adjust the transmission power of the mobile to compensate for fading. The problem with open-loop power control is that the reverse link is not identical to the forward link due to multipath fading. As a result, a faster closed-loop power control is used to combat multipath fading. The outer loop tracks frame error rates to indirectly measure the impact of multipath fading and the inner loop involves 1 bit feedback from the base station to the mobile based on the measured SIR values [23]. In current 3G WCDMA proposals, users transmit a pilot signal on both the reverse link and the forward link to enable the base station to directly measure received power from the pilot signal. When the mobile receives the power command from the base station to update transmission power, the step size can be fixed or adaptive. It is shown in [24] that the adaptive step size algorithm is superior to the fixed one because the former is faster to get to the desired power level while the latter is easier to implement.

While each power control technique has some disadvantages as indicated above, different approaches of integrating power control with other techniques are proposed in the literature, such as integrated power control and base station assignment [25], joint power control and beamforming [26], and newly proposed integrated power control and rate allocation (2.2.4) in 3G wireless CDMA systems.

2.2.3 Rate Allocation Techniques

As indicated in last section, the near-far problem exists in wireless CDMA systems. This can be mitigated by using transmitter power control. However, power control will cause the overall interference on each link (uplink) to increase due to short duration peaks in transmission power to compensate for deep fades. In cellular CDMA systems, perfect power control scheme can cause interference to increase in the neighboring cell because the users at the edge of the cells often have to use very high transmission power for the purpose of guaranteeing the QoSs, thus makes the inter-cell interference increased. Moreover, for a mobile station the path gains (path loss and shadowing) to different base stations are

independent. If power control is used to combat shadowing in one cell, it may cause additional interference in other cells. Therefore, although power control is necessary, it is not always desirable to use it alone. In addition, most of the works that have been done with the second generation wireless networks assume that the system transmits with fixed rate. However, the third generation wireless systems are designed to support multimedia services. If we still use equal rate allocation scheme, it provides fairness to all users with respect to delay, but it is power inefficient since it completely relies on transmission power control to compensate for channel degradation and results in the increase of total average required transmission power [27], which resulted in the degradation of the system capacity. This can also cause a severe problem in power restricted systems to reduce the data rate under total power constraint. Due to these many reasons listed above, we need rate allocation technique because it is an important way to allocate radio resource when packet data service is introduced in 3G wireless CDMA systems.

Consider an integrated voice and data transmission multimedia system which offers a constant bit rate voice service using power adaptation and a variable bit rate data service which can tolerate delay using rate allocation. Rate allocation ensures the required delay/jitter and the loss requirements for different connections. The motivation for applying rate adaptation to data traffic is that rate adaptation provides a power gain over power adaptation which results in a reduction of interference to other cells in a celluar structure leading to a capacity gain while still maintaining the same average data rate (or throughput) and bit error rate. CDMA schemes lend themselves to rate adaptation in a simple manner by using multiple codes, multiple processing gains, or multirate modulations [28]. In a Direct Sequence CDMA system, rate allocation can be done by changing the processing gain or by changing the number of spreading codes used for a transmission which is referred to as the multiple code scheme. The former is a single-code transmission scheme with a variable processing gain which is defined as the ratio of chip rate to user information bit rate. It allocates to each user one CDMA code channel with a processing gain that varies inversely proportional to the user information bit rate. The latter allocates multiple code channels to

each user to avoid high rate transmission. In this scheme, a high-rate data stream is first split into several fixed low-rate streams. The multiple data streams are spread by different short codes with the same processing gain (the same chip rate and data rate) and are used in parallel to achieve different data rates. Multiple short codes for one high-rate data transmission should be orthogonal over an information bit interval to reduce the intercode interference. In [29], the capacity of these two schemes are investigated. It is shown that if the maximally receivable power of a call of each class is identical in both systems, the capacities are also identical in non-fading channels. However, in multipath fading channels, the multicode system is shown to be better than the variable processing gain system in terms of capacity.

Due to the importance of rate allocation in the next generation wireless networks, lots of research has been reported in the literature. In [30], adaptive rate allocation is shown to provide larger system throughput than adaptive power control with constant data rate. Zhao and Mark [31] analyze the throughput gains for ideal rate adaptation over non-adaptation in a WCDMA system and show that rate adaptation with restrictions dose not always bring gains over non-adaptation. Hwang and Cho [32] propose a dynamic rate control method for throughput enhancement and different QoS support of each user in 3GPP WCDMA system. In [33], an adaptive transmission rate control scheme is proposed to shorten average message delay time for available bit rate services with reduced interference power. Kwok and Lau [34] suggest six efficient rate allocation schemes that are designed based on different philosophies to evaluate the effectiveness of them. The fairness and overhead problems in the rate adaptation issue of rate adaptive services in mobile networks are studied in [35]. Yang and Hanzo [36] investigate adaptive rate transmissions using variable spreading factors (variable processing gains) to increase throughput without wasting power, without imposing extra interference upon other users and without increasing the BER (bit error rate). In [37], the law that governs the dynamic data rate allocation is analyzed by adjusting target SIR thresholds to minimize total transmission power for given system throughput in CDMA systems.

Some of the works indicated above use rate allocation techniques only (with constant transmission power). While the others use rate allocation with adaptive transmission power which is referred to as integrated rate allocation and power control scheme. In the next section, we will take a look at this scheme which has attracted extensive research work recently.

2.2.4 Integrated Power Control and Rate Allocation

A typical 3G wireless communication systems should be able to support voice, data, video services, *etc.* For the sake of theoretical research, we usually assume an integrated voice and data transmission system. Voice represents the constant bit rate service which is delay sensitive but can tolerate some loss. For this type of service, power control with constant data rate is enough to meet the QoS requirement. On the other hand, data transmission represents the variable bit rate service which is loss sensitive but can tolerate some delay. For this type of service, we must use rate allocation to provide the power gain which results in the increase of system throughput. That is, under good channel condition higher data rate can be provided to increase throughput while under poor channel condition data transmission rate can be temporarily lowered even to zero to guarantee voice quality and to save transmission power which leads to a capacity gain. Therefore, for an integrated system, we employ integrated rate allocation and power control techniques. While it has been shown in [30] that rate allocation provides larger throughput than adaptive power control, combined power control and rate allocation can further improve the system capacity and throughput which is proven in [38].

We can also find numerous papers regarding this integration technique which are generally based on different optimization criteria. Authors of [39] propose an adaptive rate and power control strategy to maximize the total weighted throughput. In general, [28] and [40]-[45] develop the integrated rate and power control schemes to optimize the total system throughput which is the most frequently used optimization criterion. However, there are different methods to achieve this goal. For example, [28] obtains an upper bound

to the maximum throughput and suboptimal low-complexity schemes are considered. [40] proposes a simple heuristic rate allocation scheme, greedy rate packing, applicable in both uplink and downlink directions. [41] uses two sub-optimal rate and power control schemes, named iterative water-filling and iterative constant-power transmission to fulfill the criterion with no coordination while reducing complexity. [42] equals this optimization problem to the problem of maximizing information rate. [43] enhances the system throughput by using closed-loop power control in conjunction with rate allocation in the presence of Doppler effect and delay in feeding back channel state information. [44] solves the throughput maximization problem using a nonlinear programming approach. [45] formulates the throughput maximization problem as a classical optimization problem, modeling the constraints arising from the data rate requirements and power budgets, *etc.* Some works [46, 47] formulate the integrated power control and rate allocation to satisfy certain QoS requirements. These QoS requirements can be the average sum of data rates, minimum sum of transmission power and data rates, signal-to-interference-plus-noise ratio, a minimum rate guaranteed to each user, BER, signal-to-noise ratio, delay time, signal-to-interference ratio, minimum mean transmission delay, outage probability (probability of the data rate being less than a tolerable threshold), *etc.* Duan *et al.* [48] propose power and rate joint allocation algorithms based on the utility function which is an increasing function of a mobile user's effective throughput. Their optimization problem is to maximize total utility of the system. There are also many other different works with different methods satisfying different optimization criteria.

In order to support more users and provide higher quality services, we need to efficiently allocate scarce resources such as power and rate. Integrated rate allocation and power control techniques are very popular and effective in 3G wireless CDMA systems. However, there is limit for every technique or every combination. Therefore, if we want to further increase the system capacity and reasonably allocate resources, some new techniques or new combinations are required which have already been discovered and studied by researchers. In [49] and [50], the authors apply fuzzy logic concept to adaptive power control and rate allocation. Leelahakriengkrai and Agrawal [51] investigate the use of power control, rate

allocation and scheduling in CDMA systems to accommodate diverse service requirements. With congestion control, all mobiles can also control their transmission power and rate simultaneously in [52]. Combined rate/power, time and sector allocation is proposed in [53]. Anpalagan and Sousa [54] develop a novel combined rate, power and cell control scheme. In [13], the authors combine power control, rate allocation and base station assignment with controlled handoff switchings and dropped calls in a multi-cell environment. Zhao, Shen and Mark [55] have proposed a resource management scheme which comprises a combination of power distribution, rate allocation, service scheduling and connection admission control, jointly considering the physical, link and network layer characteristics. Liang, Chin and Liu [56] present a joint downlink beamforming, power control and rate allocation technique suitable for DS-CDMA systems with multimedia services. However, we should be aware that there always exists a tradeoff between capacity improvement and complexity.

2.2.5 The Proposed Power Control and Rate Allocation for Link Layer Scheduling

In our work, power control and rate allocation are used to provide the information such as required transmission power level and rate needed by upper layers (*i.e.*, the link layer). We present the proposed power control/rate allocation technique in this chapter and leave the involvement with link-layer scheduling to the next chapter.

Fast Closed-Loop Power Control

Fast closed-loop power control (CLPC) is applied to our scheme in the downlink to optimize the system performance. The merit of CLPC has two aspects compared to the alternative open-loop power control (OLPC):

(1) Under relatively good channel conditions where fast fading is not severe, the transmission power for mobiles from the base station can be kept to the minimum required level to satisfy the SIR at all times since CLPC performs faster than channel fading rate. Thus, CLPC is able to compensate medium to fast fading and inaccuracies in OLPC [15]. As a result, more transmission power of the base station remains for voice users that are either

far away from the base station or in a difficult environment, and data users, giving rise to enhanced system capacity and data throughput in multimedia networks.

(2) Under poor channel conditions where users undergo severe fast fading, OLPC may fail to adapt to the required transmission power for each mobile due to the slow rate and inaccuracy of OLPC, resulting in an insufficient power level to combat channel fading and fulfill QoS requirement. Whereas our approach CLPC solves this problem because the transmission power is always adjusted accordingly to satisfy the QoS requirement.

VSG Rate Allocation

Single-code variable spreading gain (VSG) rate allocation technique is applied to our scheme, as demonstrated later, to adjust the transmission rate when the satisfaction cannot be achieved by power adaptation only. According to the calculation of the transmission rate R after spreading, $R = W/\Gamma$ where W and Γ represent the spread spectrum bandwidth and the spreading gain, respectively, the actual transmission rate that is obtained can be adjusted by different value of Γ, given the fact that W is a constant. Typically, in the WCDMA standard, $\Gamma \in \{4, 8, 16, 32, 64, 128, 256, 512\}$. The actual value assigned to Γ depends on the transmission rate demanded. Note that R is inversely proportional to Γ. When we select a smaller Γ, we will obtain a larger R, and vice versa.

Integrated Power Control and Rate Allocation

The way that we combine power and rate for our scheduling scheme is to first adjust power with a fixed rate to the point that the change of power no longer produces any effect on the scheme. Then the power is fixed and the spreading gain will be adjusted to obtain the satisfactory transmission rate and to fulfill the predefined goal. More details on this integration and the interaction with the proposed link-layer scheduling scheme are expected in the immediately following chapter.

Chapter 3

Link Layer: Packet Scheduling

Third-generation wireless cellular CDMA is one of the favorable future wireless mobile network techniques. It integrates hybrid traffic and supports diversified Quality-of-Service (QoS) requirements of voice and data. However, the well-known interference-dominant characteristic imposes constraint on the capacity of CDMA systems. At downlink direction, although the effect of interference is relatively weaker than that at uplink direction primarily due to a fixed other-cell interference level, base stations' limited total power defines a stringent restriction on available downlink resources. Resource allocation techniques such as power distribution and rate allocation have been deployed to efficiently make use of scarce physical resources (*i.e.*, power, rate). Numerous allocation algorithms [17, 57] are found in the literature as a result. Recently, cross-layering issues have been attracting increasing attention. Higher-layer resource management algorithms [13, 55] are developed on top of physical layer to manage power/rate allocation. Hence, there emerges the problem of integration and interaction among different layers to better utilize physical layer resources (power/rate) and to optimize overall system performance.

3.1 Introduction

Packet scheduling is a promising link layer management mechanism. More and more research [58, 59] on this technique has emerged. However, existing algorithms focus on data scheduling with little or no concern on voice. The time-varying nature of wireless channels

determines the importance of voice scheduling. Voice dropping is less likely to come from congestion than data dropping due to its higher priority. Nevertheless, voice traffic is prone to suffer a deep-fading channel, and consumes a large amount of limited base station power, thus must be scheduled properly. In the first part of this chapter, we propose a novel voice scheduling scheme named Modified Adaptive Priority Queuing (MAPQ) for forward link in a wireless cellular WCDMA environment. An adaptive priority profile is defined in the scheme based on queuing delay and required transmission power, which borrows the idea of composite metric imported by IGRP (Interior Gateway Routing Protocol) and EIGRP (Enhanced Interior Gateway Routing Protocol) routing protocols from wired networks. This profile definition ensures system capacity improvement, packet dropping probability reduction, and fairness, testified by numerical results. The optimization of the proposed scheme are of two folds:

(1) By employing fast closed-loop power control, channel fading is precisely compensated and maximum system capacity is achieved.

(2) MAPQ ensures higher system throughput, lower voice packet dropping probability, and fairness.

Details are explained in the following sections. The necessity of voice scheduling is proven by simulation results. For the second part of this chapter, we present a unified link-layer packet scheduling framework to integrate voice and data traffic management as in multimedia networks, which is motivated by the following argument:

1. Due to the fact that more and more voice and data users access wireline backbone through the same air interface, where the problem of prioritizing and allocating resources among users arises, 3G or beyond 3G CDMA systems have to enable multi-class subscribers and multimedia traffic. Both real-time and non real-time traffic must be jointly considered when optimizing system performance such as capacity/throughput enhancement, loss/dropping probability reduction, QoS satisfaction degree improvement, resource consumption efficiency, fairness, etc, in order to achieve general optimization.

2. Efficient and effective resource management and control algorithms always come at the cost of implementation complexity since they require extra network resources for abundant communication overhead. It is therefore wise to compromise on this tradeoff by employing consistent scheduling means for both classes.

3. Different natures and QoS requirements of the voice and data traffic mean that the scheduling methods for both classes cannot be exactly the same. Existing scheduling algorithms fall into one of four categories: (*a*) scheduling only one class without coordinating multi-classes (*i.e.*, schedule data users only with little or no concern on voice [59]), (*b*) coordinating multi-classes with no scheduling for single classes [55], (*c*) scheduling individual classes with completely distinct schemes [58], resulting in implementation complexity at base stations, and (*d*) scheduling individual classes with absolutely the same scheme [60], failing to exploit distinguishing performance optimizing manners of the two classes.

We reserve the advantages and remove the disadvantages of the above algorithms by proposing an integrated framework to conform voice/data scheduling. The performance gains of the proposed framework is demonstrated through simulation. Details are explained in the subsequential sections.

3.2 Link Layer Model

The E_b/I_0 used here is expressed as:

$$\gamma_i = \Gamma \frac{P_i G_{bi}}{(P_T - P_i)G_{bi} + \sum_{j \neq b} P_T G_{ji}}, \tag{3.1}$$

where $\Gamma = W/R$ is the spreading gain with R, $R \in \{R_v, R_d\}$, the required transmission rate of voice and data packets. The actual received SIR for each MS takes the form of (2.2), where Γ is defined the same as the above.

For link-level scheduling implementation, each BS has a scheduler to regulate incoming hybrid traffic, and arrange packets aimed for MSs in a specified sequence depending on the

logic of the scheduler. The scheduler implemented in BSs is sketched in Fig. 3.1, where n_v and n_d represent the number of voice and data queues, respectively.

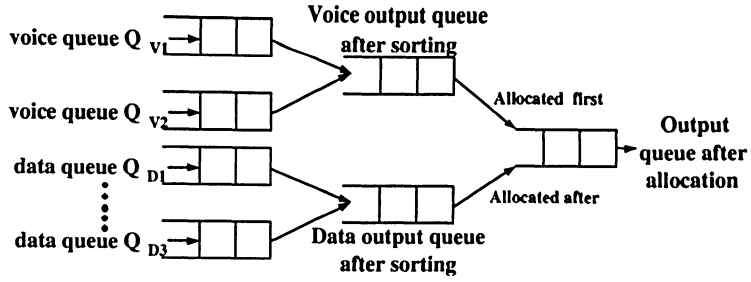

Figure 3.1: Base Station Scheduler Structure for Scheduling Framework.

3.3 MAPQ-Voice Only

In general, MAPQ has two sub-processes: sorting and allocation. The operation of MAPQ scheme and its sub-processes is described as follows:

MAPQ scheduler sorts incoming traffic into high (Q_{V1}) and low (Q_{V2}) priority queues based on each packet's calculated priority value, which is evaluated by jointly considering required power and buffering delay as $AP_i = a * delay_i^{nor} + b/power_i^{nor}$, where AP_i denotes the adaptive priority for packet i. $delay_i^{nor}$ and $power_i^{nor}$ are the normalized buffering delay and the normalized required power of packet i, respectively. Let $delay_i$, V_{thre} denote the buffering delay of packet i, the delay threshold of voice beyond which voice packets are put into Q_{V1}. Let $power_i$ and $power_{mean}$ denote the required transmission power of packet i, and the mean downlink transmission power of active users in one cell. Then we have $delay_i^{nor} = delay_i/V_{thre}$, and $power_i^{nor} = power_i/power_{mean}$. The way to normalize the delay and power components in the AP expression ensures the two terms in AP ($a * delay_i^{nor}$, $b/power_i^{nor}$) comparable. The parameters a and b are the adaptive factors determining the weight of delay over power (wdp, In our case voice is delay-sensitive thus should be assigned a larger wdp). The smaller required power (the better channel condition), or the larger delay yields the higher priority (AP). The way to define the priority profile ensures users with

better channels and larger delays to get serviced first, thus increases the throughput, reduces the voice packet dropping probability and guarantees the fairness which is measured by the mean delay in the network.

MAPQ scheduler then allocates resources starting from Q_{V1} according to AP values and total transmission power budget of BSs. The scheduler does not terminate even if the user currently being serviced in Q_{V1} requires a power exceeding the remaining budget. Since each user's priority depends on both delay and power, higher priorities do not merely indicate smaller powers. Also, after Q_{V1} has been fully checked and if there is power remaining (P_r), the scheduler will continue to check Q_{V2} and service available users in Q_{V2} using P_r. This is the major difference between MAPQ and classic PQ. A similar concept can be found in [60], where the authors proposed a modified FIFO scheme for power-constraint systems. This modified queuing fashion can lead to higher power utilization efficiency, as will be shown later.

The detailed implementation steps of this scheme are illustrated in Fig. 3.2 (sorting sub-scheme) and Fig. 3.3 (allocation sub-scheme). Note that the power budget *pwt* in Fig. 3.3 is initialized with the downlink traffic power of the base station (in our case $70\%P_T$, see Section 3.5.2).

3.4 UF-Unified Voice/Data Scheduling Framework

We demonstrate the easy integration and general adaptability of MAPQ to further apply it to the UF. Compared to countless current research work, the proposed framework has several features.

Firstly, voice and data scheduling algorithms in the proposed framework have coherent infrastructure, defined by the proposed Adaptive Priority Profile based on delay/power(/rate for data only) in a composite fashion. It renders hybrid voice/data traffic robust to variable combination ratio in terms of some performance criteria which should be held stable and optimal at all times, such as downlink power efficiency.

Secondly, the framework also allows dissimilarity between the scheduling of voice and

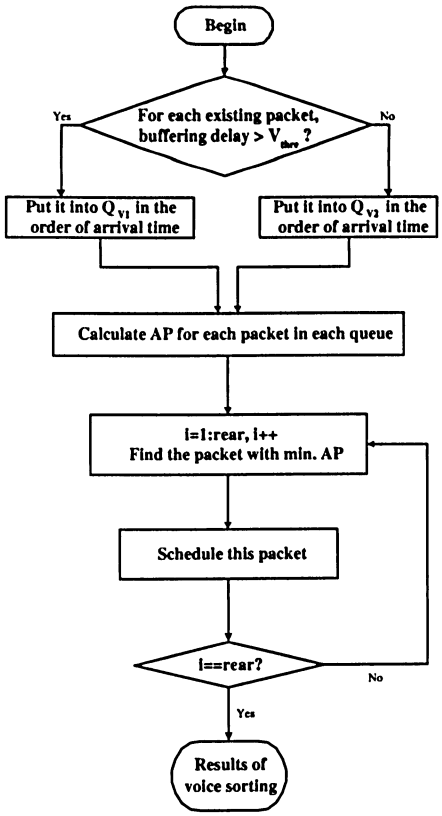

Figure 3.2: MAPQ Scheme–Sorting Sub-process.

data in that these two types of traffic have different nature (voice has constant bit-rate and is delay-sensitive, while data requires variable bit-rate and is delay-tolerant) and QoS requirement (*i.e.*, different γ^*). Although in this section we focus on addressing a framework for integrated traffic and seek for overall performance improvement, the featured performance for individual class (*i.e.*, dropping probability for voice and loss rate for data) cannot be sacrificed. Therefore, the dissimilarity is not negligible and will be discussed in Section 3.4.2.

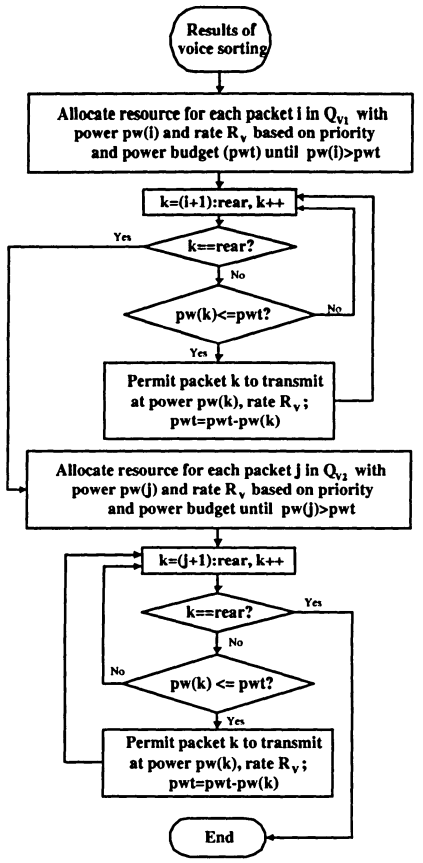

Figure 3.3: MAPQ Scheme–Allocation Sub-process.

3.4.1 Similarity of Voice and Data Scheduling

We describe the framework for scheduling voice and data as follows. For convenience, the framework is divided into two sub-processes, namely, sorting sub-process and allocation sub-process.

Sorting: For both voice and data classes, the scheduler sorts incoming traffic into different queues based on each packet's buffering (queuing) delay incurred at the base station which serves as an event trigger. It then sorts within each queue by calculating adaptive priority value for each packet of that queue, which is defined by jointly considering required power/rate and buffering delay as

$$AP_i = a * delay_i^{nor} + b/power_i^{nor} + c/rate_i^{nor}, \tag{3.2}$$

where AP_i denotes the adaptive priority for packet i, a, b, and c are adaptive constants. Voice and data have different a, b, and c values. Note that "$rate_i^{nor}$" denotes the normalized new required transmission power after decreasing the data rate, using (3.1) where $\Gamma = W/R_d$. The normalization method is similar to that used for voice power normalization. Here we use "$rate_i^{nor}$" in order to distinguish from $power_i^{nor}$ in the AP expression. Furthermore, "$rate_i^{nor}$" is used to emphasize that the new required transmission power is obtained by adjusting the data rate. More details will be addressed in Section 3.4.2. The larger delay, or the smaller required power, or the lower transmission rate (thus smaller required power to get identical received γ_i) yields the higher priority (AP). Therefore the earlier the packet gets transmitted.

Allocation: The scheduler allocates power resource starting from the packet with the largest AP (the head of the queue) based on the total power budget of base stations. It does not terminate even if the user currently being serviced requires a power exceeding the remaining budget. The reason has been discussed before.

In the proposed framework, this allocation scheme is applied to not only each queue within the class but also among classes. It is apparent that voice class has higher priority than data class, which necessitates the employment of the proposed allocation algorithm (modified PQ) for each voice queue to secure the scheduler not skipping to data class before completing checkup within voice class. However, classic priority queuing scheduler will not jump to data users until all the voice users have been serviced. It may cause power waste if none of the users in voice queues but some of the users in the data queue can be serviced. The proposed allocation algorithm further improves system capacity and throughput by also performing an exhaustive search within data class after an exhaustive search within voice class.

3.4.2 Discrepancy of Voice and Data Scheduling

The necessity of differentiating voice and data scheduling has been aforesaid and is shown later in the simulation section. We will describe in detail where the difference exists in this section.

Sorting

Difference in sorting mainly lies in the expression of adaptive priority profile. We have $AP_i = a_1 * delay_i^{nor} + b_1/power_i^{nor}$ for voice and $AP_i = a_2 * delay_i^{nor} + b_2/power_i^{nor} + c_2/rate_i^{nor}$ for data, attributing to distinct natures of voice and data traffic. Voice requires constant bit rate during transmission and thus it is unlikely to change voice users' priority through rate variation. On the other hand, data (best-effort traffic in this paper unless otherwise specified) transmission rate is variable and can be raised if extra power is available or reduced without enough power resource to support target rate. In data profile expression, the last term "$c/rate_i$" is not used until delay exceeds a pre-defined threshold D_{thre}, and the required transmission power is still too large to increase AP. This is the only case where we decrease the required transmission data rate in order to get a smaller required power while SIR target is maintained.

Moreover, the ratio of a over b (or c) determines wdp. The parameters a, b and c here are adaptive factors through which wdp for voice and data is adjusted. In general, voice is delay-sensitive and will be dropped if the delay exceeds a maximum tolerable value (*i.e.*, $100ms$). As a result, its wdp is larger than 1, translating into a larger weight of delay to reduce voice packet dropping probability [61]. In contrast, data users are able to withstand some delay and do not have strict drop bound, in which case wdp should be less than 1 to service users under desirable channel conditions (small transmission power) with preference.

For simplicity, we set a to 1 and adjust b in the range of $(0, 1)$ for voice. For data, a is set in the range of $(0, 1)$ while b (or c) is set to 1.

The subsequent description elaborates on data sorting sub-process as an instance and reveals multi-performance enhancement gained by the adaptive priority profile definition.

Data packets destined to the mobile users are sorted in the base stations into one of the three queues based on the packets' buffering delay. We design these data queues and the associated sorting steps as follows.

Queue "normal" (Q_{D1}): $delay = 0$, $AP = b/power$, $(b = 1)$. This is the initial case or right after a packet is transmitted. Users experiencing better channels get serviced first, giving rise to better data throughput.

Queue "moderate" (Q_{D2}): $0 < delay <= D_{thresh}$, $AP = a * delay + b/power$, $(b = 1, c = 0)$, where D_{thresh} denotes delay threshold for data packets. This is an interim corresponding to when delay is still tolerable. AP is increased mildly with $a = 0.5$ in our simulation to transmit packets undergoing moderate channel fading in order to avoid excessive delay causing congestion and loss.

Queue "urgent" (Q_{D3}): $delay > D_{thresh}$, $AP = a * delay + c/rate$, $(b = 0, c = 1)$. In this aggressive stage, we have to decrease data transmission rate as well as set a to a larger value (in the simulation, 0.7) to increase AP and service users in very poor channel conditions, contributing to fairness fulfillment.

This sorting process is demonstrated in Fig. 3.4.

Although the voice sorting sub-process employs two queues as discussed earlier, voice queues and the associated sorting steps can be interpreted similarly to the steps mentioned above for data: $0 < delay <= V_{thre}$, $AP = b/power$, $(a = 0)$ for *Queue "normal"* (Q_{V2}), and $delay > V_{thre}$, $AP = a * delay + b/power$, $(a = 1, b = 0.8)$ for *Queue "urgent"* (Q_{V1}), where V_{thre} denotes delay threshold for voice packets.

Allocation and Reallocation

The allocation sub-process is the same for scheduling voice/data up to this point. Nevertheless, data scheduler does not end in here as voice scheduler does, because data transmission rate is adjustable according to the amount of remaining power resource as indicated earlier. Hence, the remaining power (if any) can be further allocated to specially selected data users with various transmission rate values. The selection procedure is demonstrated by the

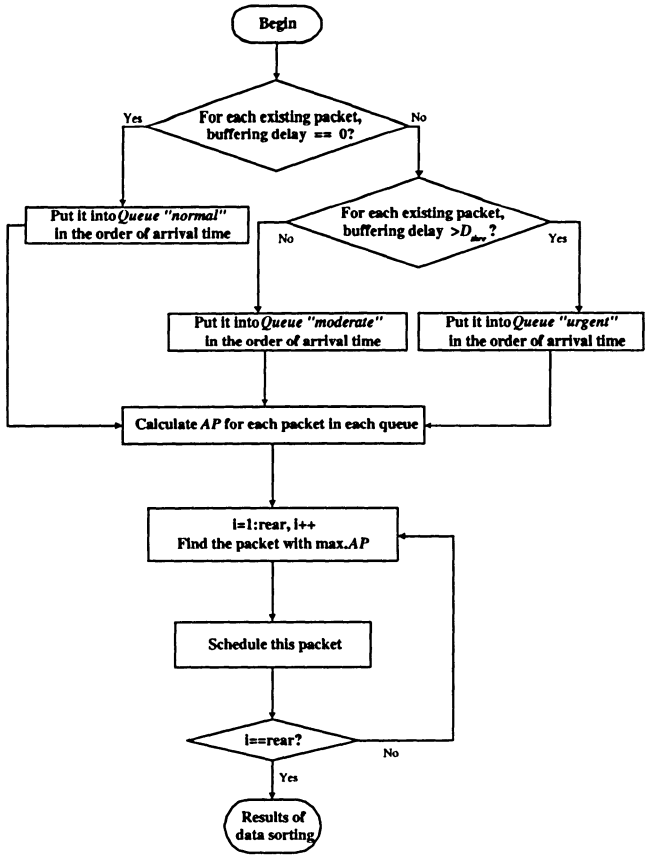

Figure 3.4: Sorting Sub-Process of Data Scheduling.

following algorithm. We continue to design a branch named reallocation of data allocation sub-algorithm as:

1. If there is power remaining, check if all the "normal" packets ($AP = b/power$) are served. If yes, use remaining power (P_r) to allocate equal rate among the "normal" packets.

2. If no, check if P_r is enough for the minimum transmission rate of the first "normal" packet still waiting in the queue. If yes, serve this packet with the maximum rate allowed by P_r.

3. If no, check if there has been any "normal" packets serviced. If yes, share P_r with

equal rate among them.

4. If no, allocate P_r to the "moderate" or "urgent" packet which has largest AP and has already been allocated resource.

The detailed implementation steps of this process is illustrated in Fig. 3.5. Note that in data scheduling, *Queues "normal", "moderate", "urgent"* (Fig. 3.4) only represent abstract priority profiles according to which a user's priority is calculated, depending on its buffering delay (event trigger). These queues are in fact conceptual and do not exist physically. All the data packets enter one physical buffer, no matter what *Queue* it belongs to, and the scheduler is programmed to perform the processes described above. The "queue" that we mention in the first step of Fig. 3.5 denotes the physical queue, hence is not distinguished among the three logic queues proposed. While in voice scheduling, high and low priority queues Q_{V1} and Q_{V2} (Fig. 3.2) physically exist.

The idea behind is summarized as follows. After every packet in the queue has been checked, if there is remaining power for one more "normal" packet even at minimum rate, this packet gets serviced (system capacity increased). If not, share the power left among "normal" packets since they require lesser power (data throughput improved). Only if neither of the above is true, we give the extra power to the first "moderate" or "urgent" packet in the sorted output queue. This packet has already been allocated resource after all packets in the data queue have been scheduled. P_r is not allotted to "moderate" or "urgent" packets which are still in the queue since they consume relatively large powers. As queuing delay increases, they will be assigned a larger AP until eventually get transmitted. This is the key role of delay taken into account in our design.

3.5 Simulation Senario and Results

3.5.1 Simulation Senario

Simulation is carried out in a cellular WCDMA environment with chip rate $3.84Mcps$ and spread spectrum bandwidth $5MHz$. Fast closed-loop power control is performed for every

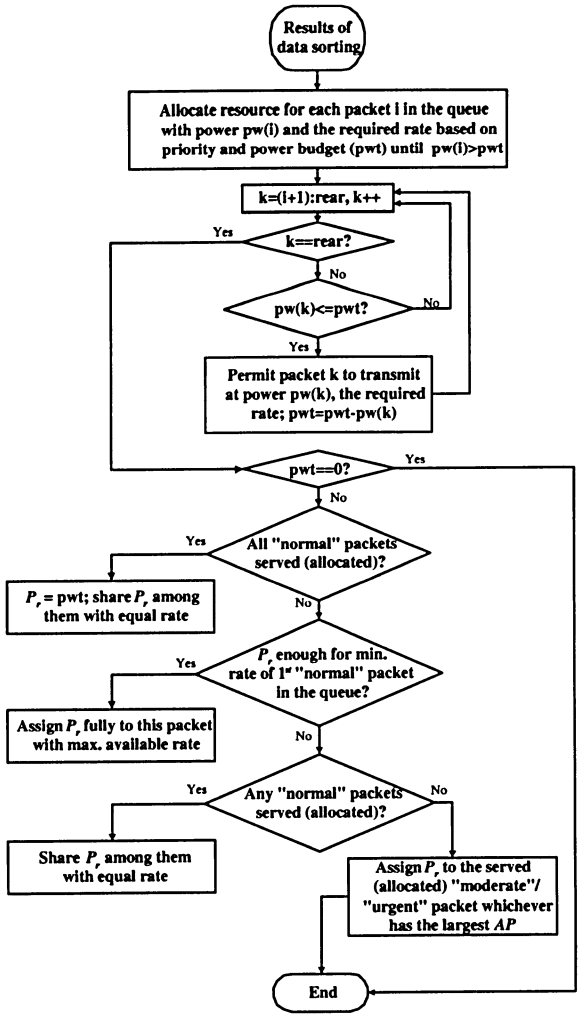

Figure 3.5: Allocation and Reallocation Process of Data Scheduling.

slot of time (approximately $0.667ms$) in each frame (consisting of 15 time slots) with a fixed step size of $1dB$.

Other relevant parameters are: 19 wrap-around cells with radius $r = 500m$ (macro-cell). One BS is located in the center of each cell with $P_T = 20W$ and a portion of 70% of P_T is dedicated to traffic channel [62]. Mobility speed in Rayleigh fading model is $10km/h$ (vehicular environment), $\alpha = 4$, $\sigma_X = 8dB$, $\sigma_Y = 2dB$, $\gamma^*_{voice} = 5dB$, $\gamma^*_{data} = 3dB$.

Hybrid voice and data users are uniformly distributed, approximately 30 users per cell on average. Voice traffic is modeled as "ON-OFF" with 50% "ON" duration probability, and best-effort data traffic is generated with exponentially distributed arrival rate. Generally speaking, voice traffic has lower transmission rate compared to data traffic. In the integrated voice/data scheduling scheme, minimum voice rate R_v is selected from one of the following values: $\{8, 16, 32, 64\}kbps$ corresponding to a spreading gain of 512, 256, 128, 64, respectively, while $R_v = 64kbps$ is the fixed transmission rate in the voice-only scheduling (MAPQ). Data rate R_d can be chosen from any available value allowed by the spreading gain set of $\{4, 8, 16, 32, 64, 128, 256, 512\}$.

For voice-only scheduling, maximum tolerable delay is $d_{max} = 100ms$, buffering delay ($d_b = 60ms$) is used in the simulations to determine the unfairness criterion. The power for calculating AP is procured from (3.1). For integrate voice/data scheduling framework, the delay thresholds for sorting voice and data packets into different queues are $V_{thre} = 10ms$, and $D_{thre} = 120ms$. The delay bounds of voice and data packets are $100ms$ and $2s$, respectively, which will be used in the mean normalized delay calculation in the following section.

3.5.2 Numerical Results

We first define the performance measures used in the MAPQ and UF simulations for voice and data. Define N_F, N_A, N_S,N_C, ψ_F, ψ_A as the total number of users in the network (in our case 30/cell), the number of active users in the network, the number of active users actually serviced, the number of cells in the system (in our case 19), the throughput ($kbps$) of the system if all the active users can be serviced, the actual throughput of the system, respectively.

1. MAPQ for voice only:

    - Normalized system capacity (throughput)-ψ_A/ψ_F. Note that voice packets have constant transmission rate thus the capacity and throughput have similar behaviors.

- Voice dropping probability-(number of packets dropped)/(number of packets transmitted). A voice packet is dropped if its buffering delay exceeds the delay bound ($100ms$).
- Unfairness probability-we call it "unfair" if a user's buffering delay is greater than d_b yet not serviced. Therefore, unfairness probability refers to the probability that such unfair event happens. One possible way is to use N_{UP}/N_A to measure it, where N_{UP} is the number of users that experience unfairness.
- Traffic load-N_A/N_F.

2. UF for hybrid traffic:
   - System capacity-N_S/N_C.
   - Traffic throughput-ψ_A/N_C.
   - Outage probability-fraction of time that a user's received power is below the minimum acceptable power level to satisfy the target SIR.
   - Average power utilization (efficiency)-(total power consumed)/(total traffic power budget of BSs). It acts as the indicator of resource consumption efficiency.
   - Voice ratio-the proportion of voice traffic in the hybrid traffic. It controls the variation of the hybrid traffic structure.

3. For both MAPQ and UF:
   - Mean normalized delay-$\frac{\sum_{i=1}^{N_s}(normalized\_delay\_of\_packet_i)}{N_s}$, where the numerator equals to $\frac{buffering\_delay\_of\_packet_i}{delay\_bound\_of\_packet_i}$. We measure the normalized delay only for successfully serviced users because there are other criteria, voice dropping probability and data outage, to illustrate the behavior of each scheduling scheme with service failures.

Voice Only–MAPQ

Compared to systems where no sorting scheme nor modified PQ allocation scheme (FIFO) is deployed, system performances of the MAPQ scheme in terms of system capacity, voice packet dropping probability, mean normalized delay, and unfairness probability improve in various degrees, as shown in Figs. 3.6, 3.7, 3.8 and 3.9, respectively.

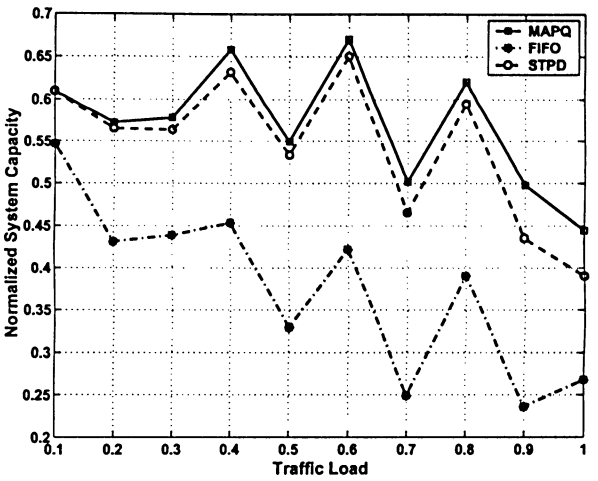

Figure 3.6: System Capacity Comparison for MAPQ.

As the traffic load grows heavier, the performance gain of the proposed scheme becomes more apparent. This can be explained by the following observations. When the power budget is getting tighter and users are more competitive for limited resources, the FIFO scheme is not capable of producing satisfactory results due to its inadaptability to severe system environment. While the MAPQ scheme is generally stable and insensitive to throughout traffic variation, and able to produce acceptable outcomes even if experiencing stringent conditions, as a result of well designed adaptive features.

Moreover, we compare the MAPQ scheme with a more advanced scheduling scheme in the literature named STPD (Scheduling with Transmission Power and Delay) [58]. In this scheme, packets whose required transmission power is less than a threshold P_{th} are classified into Group 1, otherwise are classified into Group 2. For real-time traffic like voice, if the

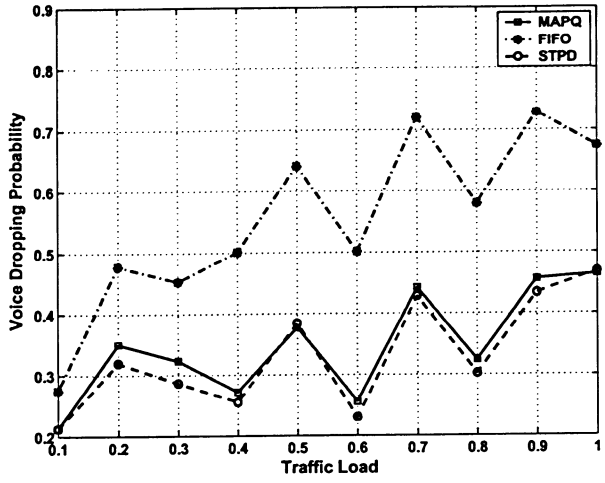

Figure 3.7: Dropping Probability Comparison for MAPQ.

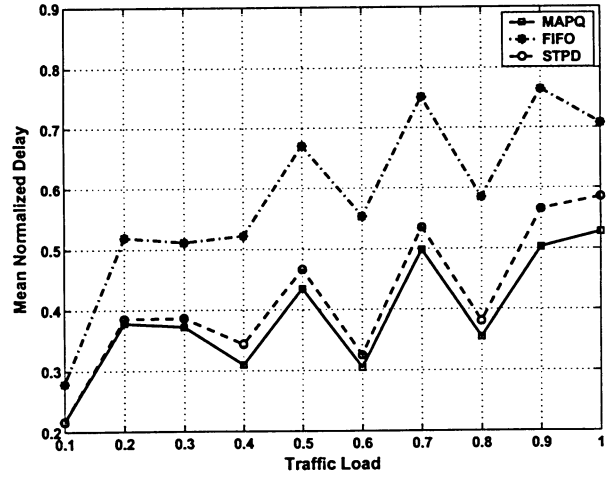

Figure 3.8: Mean Normalized Delay Comparison for MAPQ.

maximum buffering delay of Group 1 is less than a delay threshold, Group 2 is transmitted first to avoid the exceeding of the tight delay bound. While for non real-time traffic like data, Group 1 is always transmitted first since it is delay-tolerable. This algorithm is less complex in calculation since it does not use the priority to sort each packet. However, the

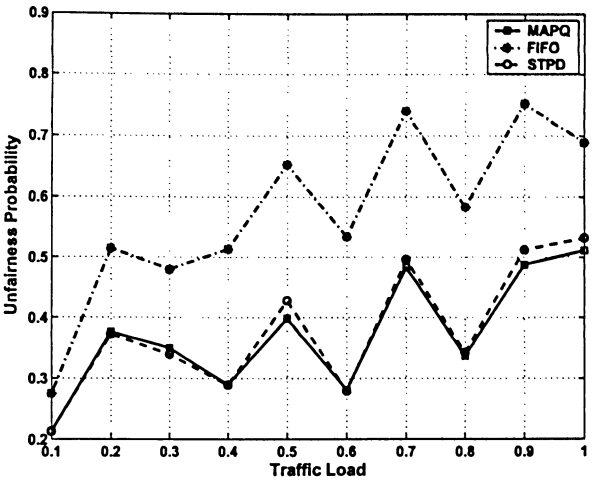

Figure 3.9: Unfairness Probability Comparison for MAPQ.

simplicity may result in some degradation of the performance, as shown in Figs. 3.6 and 3.8, where obviously the more complex MAPQ scheme performs better in terms of both the system capacity and the mean normalized delay in the network. The MAPQ scheme not only outperforms the STPD in these criteria, but also maintains other performances in terms of voice dropping probability and unfairness probability, as shown in Figs. 3.7 and 3.9. In Fig. 3.7, the MAPQ performance is only slightly worse than STPD. In Fig. 3.9, MAPQ and STPD almost have the same performance but as the traffic load becomes heavier, the MAPQ shows the trend to outperform the STPD.

Note that the fairness criterion in our simulation is implied by both the mean normalized delay and the unfairness probability measures. Smaller normalized delay and lower unfairness probability indicate higher degree of fairness.

We also testified the necessity of both sorting and the allocation sub-schemes of MAPQ scheme by comparing MAPQ with two reference cases, namely allocation (modified PQ) without sorting and sorting without allocation (classic PQ). The proposed scheme outperforms both of the references in terms of system throughput, packet dropping probability, and unfairness probability with 2%-10% performance gains (not shown in this work).

Unified Voice/Data Framework (UF)

Individual performance gain of voice under the proposed scheduling algorithm has been procured and illustrated above in terms of system capacity/throughput, packet dropping probability and unfairness probability. Note that the values of a, b, and c used in the simulation are obtained from the estimation.

In this subsection, we focus on measurable performance of hybrid voice/data traffic under the proposed unified framework. Three reference algorithms are compared with our algorithm and evaluation is realized through several significant criteria: system capacity, traffic throughput, outage probability, average power utilization, and mean normalized delay. Reference 1 employs SPS (static priority scheduling) [60] algorithm for either class, Reference 2 employs STPD (scheduling with transmission power and delay), and Reference 3 employs data scheduling only, which is the popular RF (rate fairness) algorithm as in [35, 57]. They represent Three typical simulation conditions.

Figs. 3.10-3.14 show the comparison results in terms of system capacity, traffic throughput, outage probability, power utilization efficiency, and mean normalized delay, respectively. The results are obtained by averaging over different load situations. The proposed algorithm achieves better performance in all cases in comparison with SPS and STPD. The proposed framework illustrates an appropriate combination of individual performance sustentation as well as integrated optimization and robustness. Fig. 3.10 exhibits the capability of the proposed algorithm to explore the bursty nature of the data traffic with a visible steep slope of the solid square-marked curve under the variation of traffic ratio. This capability benefits from discrepancy of voice/data scheduling (*i.e.*, different adaptive priority profiles, data reallocation mechanism) in the framework. While SPS (dashed asterisk-marked curve) appears mildest under the traffic structure change, as a result of limited capacity and resource utilization capability. Solid square-marked curves of Figs. 3.10 and 3.11 confirm the results in [63] in which system throughput (Fig. 3.11) decreases with the reduction of data portion, probably because of the fact that data traffic has much more burst and much higher bit rate, thus is more susceptible to load change and affects more on throughput and capacity (Fig.

3.10) behaviors.

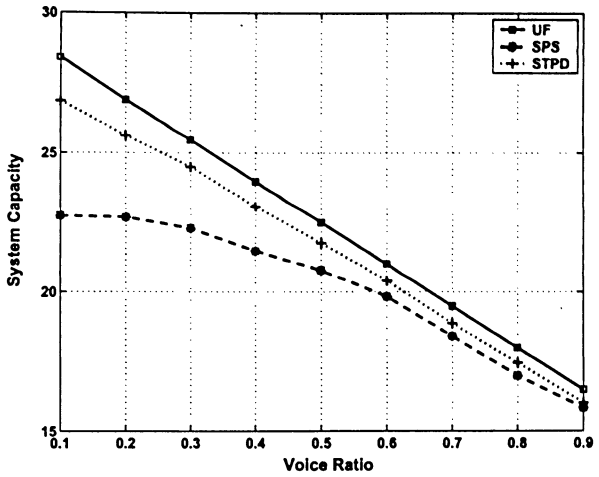

Figure 3.10: System Capacity Comparison for UF.

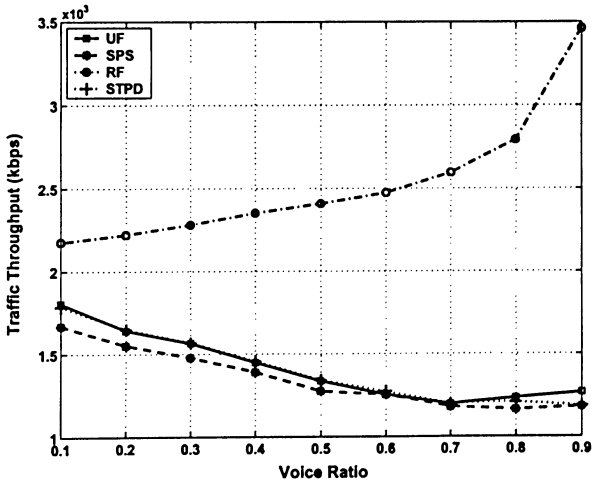

Figure 3.11: Traffic Throughput Comparison for UF.

Although RF has much higher traffic throughput (observed in Fig. 3.11, circle-marked dash-dot curve) because it distributes rate equally among all active users regardless of actual channel impairments, which comes at the expenses of users being serviced with a power below the minimum acceptable target at most of the time (circle-marked dash-dot curve in Fig.

3.12).

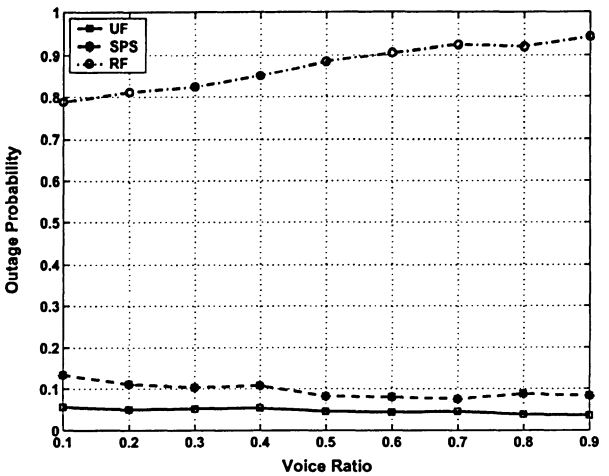

Figure 3.12: Outage Probability Comparison for UF.

This insufficient transmission power fails to combat channel fading and causes transmission failure (SIR target unsatisfied). While the power baseline in SPS is assured, because of the lack of sorting process, users are serviced in a first-in-first-out mode where a user occupying large power is possible to be allocated before users requesting smaller power. Also, due to our allocation sub-algorithm which is based on an exhaustive search mechanism, system capacity and traffic throughput of the proposed algorithm outperform those of SPS as displayed in Figs. 3.10, 3.11 (solid vs. dashed curves), respectively, without sacrificing QoS satisfaction degree (here defined by outage probability as in Fig. 3.12, solid vs. dashed curve).

Aside from the exploitation and maintenance of individual behavior features, the capability of attaining overall optimization and robustness to various traffic structure, resulting from the consistent infrastructure of the framework, is depicted in Figs. 3.12, 3.13 and 3.14, since in these figures the fluctuation of the traffic stucture does not affect UF much.

Voice traffic has less burst, smaller range and lower data rate than data traffic. Wang *et al.* [64] advocates that with the same total offered load, a larger fraction of voice permits better multiplexing and hence more efficient resource usage, giving rise to the fluctuation of

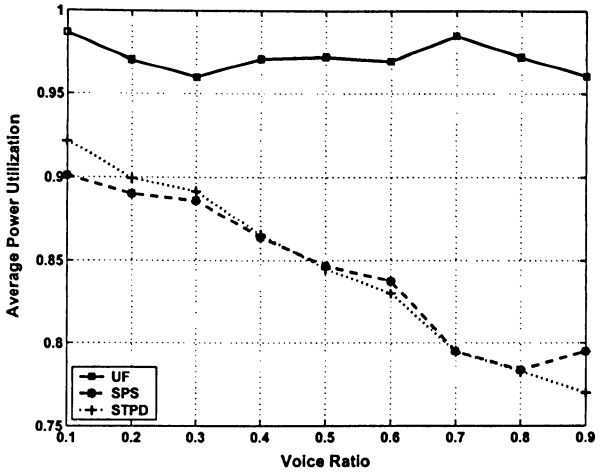

Figure 3.13: Power Utilization Efficiency Evaluation for UF.

average power utility with voice ratio, shown in their simulation results. On the contrary, Fig. 3.13 convincingly illustrates that the UF guarantees power utilization efficiency at above 97% at all times benefitting from data reallocation mechanism and is resistible to a variety of offered traffic load and voice/data ratio. At the same time, the obvious fluctuation of the dashed curve and the dotted curve reflects the vulnerability of SPS and STPD under the altered traffic structure.

Furthermore, the reallocation mechanism ensures "good" users to be better and "poor" users to be serviced, leading to fairness protection as shown in Fig. 3.14, low and insensitive outage occurrence (solid curve in Fig. 3.12, reflecting high degree of QoS satisfaction) throughout varied voice ratio.

Fig. 3.14 shows the outstanding delay performance of the UF scheme. In general, higher voice ratio yields better delay performance since data packets may experience much longer delay thus have larger impact on the delay behavior in the network. In fact, the delay behavior of the MAPQ scheme in Fig. 3.8 does not show such striking performance enhancement. We mentioned before that data packet delay is usually much larger than voice packet delay. The data packet delay performance determines the delay performance of the entire network. Therefore, we can conclude that the result in Fig. 3.14 suggests that the

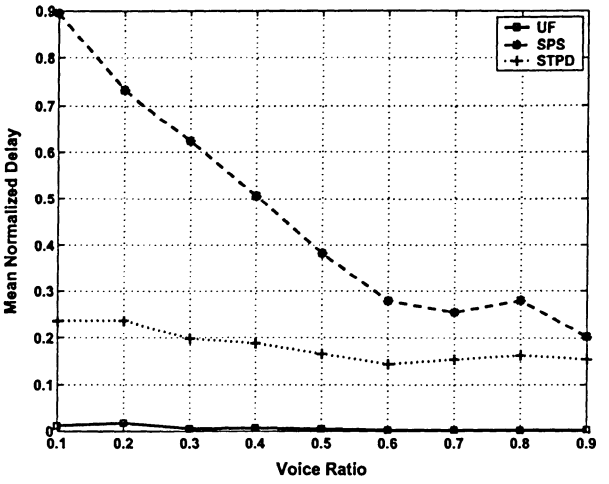

Figure 3.14: Mean Normalized Delay (Fairness) Evaluation for UF.

gain of the UF scheme come largely from the desirable delay performance of data users. Due to the dominant characteristics of data packet delay in the network, this performance gain also compensates for the voice packet delay performance, observed as a nearly flat curve (the solid square-marked curve) in Fig. 3.14. This non-fluctuated delay performance curve further demonstrates the fairness assurance among data users in the UF scheme. In addition, it implies that in the MAPQ scheme, the improved performance of voice packets does not necessarily sacrifice the data packets behavior (*i.e.*, data packet delay).

We believe all of the above gains are acquired from the unity of proposed framework and innovative sorting and allocation (reallocation) mechanisms deployed within.

Chapter 4

Network Layer: Admission Control and Soft Handoff

Next-Generation cellular CDMA networks must be able to support a heavy load of mobile subscribers and untethered wireless personal communication. High mobility and universal access are enabled by handoff mechanisms employed in such networks. Without handoff, forced termination would occur frequently as mobile users traverse cell boundaries. 2G systems such as GSM adopt hard handoff for inter-cell call transfer causing unpleasant temporary audio cutouts when handoffs take place, due to frequency reallocation in the new cell. CDMA networks such as IS-95 (2G) and next-generation CDMA systems facilitate the implementation of soft handoff because all users access the entire spread spectrum eliminating the problem of frequency reuse. In addition, it is well-known that CDMA is based on power control technique to avoid near-far effect. It necessitates the existence of soft handoff, which guarantees the mobile station to persistently connect to the base station where the strongest signal is received, for power control loop to run properly.

4.1 Introduction

It is generally impossible to have one handoff algorithm suitable for systems under different connection admission control (CAC) policies, the reason being that coordinating diverse parameters and achieving overall optimization are difficult and complicated. A detailed introduction can be found in [65] and research on selecting optimal parameters [66, 67] such

as optimum handoff coverage, active set size and add/drop thresholds, fade margin, optimum balance of new call blocking and handoff dropping probabilities, etc, has been carried out in the literature. An apparent pair of contradictive parameters representing resource management efficiency and effectiveness is the call blocking probability (P_b) and the handoff dropping probability (P_d). Since fixed downlink total traffic power is shared between newly accepted users and ongoing handoff users, one's being greedy will induce another's being starving. It is therefore important to regulate and optimize their behaviors by balancing the amount of resources distributed. Based on the fact that interrupting an ongoing call is more disagreeable than rejecting a new call, handoff users are issued higher priority to reduce P_d wherever competition arises. Among numerous prioritizing algorithms [64] [68], resource reservation has attracted overwhelming favor owing to lighter required communication overhead. However, existing algorithms prioritize handoff users as an entire category towards new users category. None of them concern priority assignment among handoff users, which is necessary when several users attempt to handoff to a same target cell, under the constraint of limited available target cell power (guard power P_G) set aside for handoffs.

We propose a novel handoff algorithm based on adaptive priority assignment for handoff users during soft handoff procedure. The priority profile is designed in a composite fashion jointly considering user mobility, user location, channel condition, and call holding time, which are essential elements to be referred in order to outline a specific user's handoff situation. Prediction is performed to acquire the information for calculating priority in advance, which is also crucial to inform among BSs of their reserved guard power in dynamic channel reservation schemes. Predicted potential handoff users with *1)* earlier handoff prediction (higher mobility, closer towards border) and faster handoff completion (longer call holding time within a predefined threshold), thus more possible resource release, and *2)* smaller required handoff power reflecting more desirable channel gain, will be prioritized based on the adaptive priority profile, in case several of them initiate handoff request to the base station simultaneously in the future. Here by "simultaneously" we mean that while the BS is still handling one handoff request, another one or several requests may emerge, and so forth over

and over. This is especially true in metropolitan cities where the mobile networks are mostly busy with cellular phone users. It does not mean two or more requests emerging at exactly the same time instant, which is not very likely in reality. Therefore, it is imperative to avoid signaling flood at the moment of simultaneous handoff requests, since predictions would not occur at the same time. The proposed adaptive prioritizing algorithm can be proven to achieve a larger handoff capacity with the same amount of resources (power efficiency) at a lower P_d, as well as higher guard power utilization.

4.2 Network Layer Model

The E_b/I_0 consider here is expressed as:

$$\gamma_i = \Gamma \frac{P_i G_{bi}}{(P_T - P_i)G_{bi} + \sum_{j \neq b} P_T G_{ji}}, \tag{4.1}$$

where $\Gamma = W/R_v$ is the spreading gain with R_v, the required transmission rate of voice packets. The actual received SIR for each MS takes the form of (2.2), where Γ is defined the same as the above.

For prioritizing handoff users, a scheduler implementing link-layer scheduling with network-layer inputs inheres in each base station to regulate incoming handoff requests, and to arrange these messages designated to different target BSs in a specified sequence depending on the logic of the scheduler, which has been proposed with physical-layer inputs in Chapter 2. The scheduler applied to BSs is sketched in Fig. 4.1, assuming the current serving base station is BS_0. BS_1, $BS_2 \cdots BS_n$ denote target handoff BSs, with n the number of handoff base station targets.

4.3 CAC and Soft Handoff

4.3.1 Connection Admission Control (CAC)

Capacity of CDMA Systems is "soft". Acceptance of each new call increases the interference level of existing ongoing calls and affects their quality. Hence, CAC is deployed to control

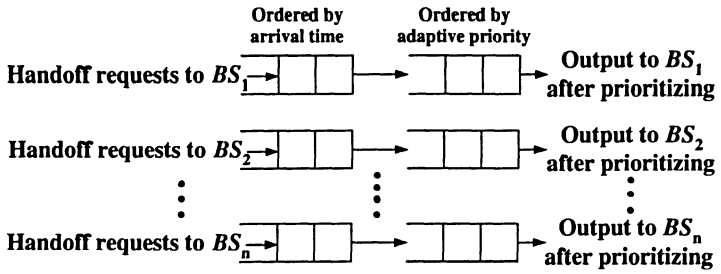

Figure 4.1: Base Station Scheduler Structure for Handoff Algorithm.

the access to such networks, complying with types of service and Quality-of-Service (QoS) requirement, as well as current system load.

With the growth of wireless Internet applications, the forward link becomes very critical to system capacity, in that the bottleneck-like power capacity of BSs imposes stringency on available power resources allotted to each sharing user. The CAC mechanism employed in this paper, thereby, is based on total downlink traffic power and the priority of handoffs over new calls. A handoff request is admitted if

$$\sum_{i=1}^{N_{on}^c} P_i + P_{ho} \leq P_t, \tag{4.2}$$

and a new connection request is admitted if

$$\sum_{i=1}^{N_{on}^c} P_i + P_{rsv} + P_{new} \leq P_t, \tag{4.3}$$

where N_{on}^c is the number of ongoing calls in cell c, P_i, P_{ho}, P_{rsv}, P_{new}, and P_t represent the required powers of an ongoing call i, the incoming handoff call, the reservation for future handoffs (will be discussed later), the incoming new call, and the downlink total traffic power of BSs, respectively. Both (4.2) and (4.3) conform to the general admission criterion $\sum_{i=1}^{N_{on}^c} P_i \leq P_t$.

4.3.2 Soft Handoff

One of the major benefits of a CDMA system is the ability of a mobile to communicate with more than one base station at a time during the call [69]. This functionality allows the

CDMA network to perform soft handoff. In soft handoff a controlling primary base station coordinates with other base stations as they are added or deleted for the call. This allows the base stations to receive/transmit voice packets with a single mobile for a single call.

In forward link handoff procedure, a mobile receives pilots from all the BSs in the active set through associated traffic channels. All these channels carry the same traffic (with the exception of power control sub-channel [69]), which facilitates the mobile to gain macroscopic diversity by combining power received from the channels (*i.e.*, maximal ratio combining [70]). Thus, less power is needed implying total interference lessening and system capacity raising.

A basic soft handoff algorithm typically used in 3G CDMA systems is illustrated in Fig. 4.2, with AS_Th, AS_Th_Hyst, AS_Rep_Hyst, and ΔT defined as the threshold of reporting for active set transfer, hysteresis of the former threshold, replacement hysteresis, and time to trigger, respectively. CPICH is the abbreviation of Common Pilot Channel. The events, together with the hysteresis mechanism and time to trigger mechanism are discussed in [71].

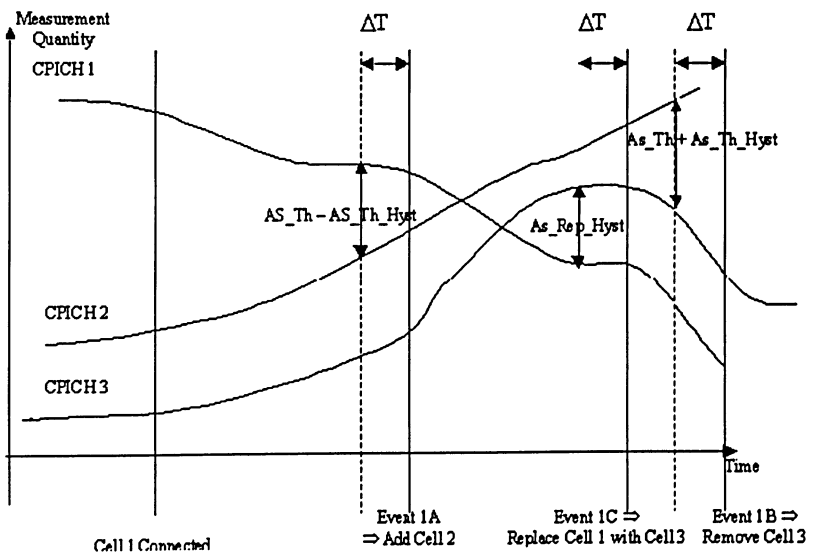

Figure 4.2: A Typical Soft Handoff Algorithm.

We employ a similar basic algorithm with slight simplification. The selection of a base station into the active set and the deletion from the active set are based on dynamic thresh-

olds. Let M_{ps}^{b} and $Best_{ps}^{active}$ be the measured pilot signal from base station b, and best measured pilot from the active set. All the variables appearing in the inequalities below have the unit of $Watt$. A base station b is added into the active set if

$$M_{ps}^{b} > Best_{ps}^{active} - AS\_Th + AS\_Th\_Hyst, \tag{4.4}$$

for a period of ΔT, and is removed from the active set if

$$M_{ps}^{b} < Best_{ps}^{active} - AS\_Th - AS\_Th\_Hyst, \tag{4.5}$$

for ΔT, where AS_Th, AS_Th_Hyst, and ΔT are design parameters.

We briefly describe the mobile-assisted soft-handoff procedure as follows: Mobile detects pilot strength from its monitored set by (4.4) and sends a Pilot Strength Measurement Message (PSMM) to the serving BS. BS requests resources from the target handoff cell, allocates traffic channel and sends a Handoff Direction Message (HDM) to Mobile. Mobile transfers this pilot to the active set and transmits to BS a Handoff Completion Message (HCM). Mobile starts handoff drop timer when the pilot strength in the active set meets (4.5) and sends to BS a PSMM. Mobile removes the pilot from the active set to the monitored set as the above time expires.

Note that the monitoring mechanism enables us to perform the prediction for prioritizing without extra network resources or high cost, as will be discussed in the next section.

4.4 Adaptive Prioritizing Soft Handoff Algorithm

The parameters and performance measures of the proposed prioritizing algorithm are addressed in this section, together with the description of the detailed implementation procedure of the algorithm. We mentioned above that the adaptive priority profile is designed by jointly considering several elements, which are critical to define a specific handoff user.

4.4.1 Prediction

First of all, user mobility and location information is needed by prediction, which is the prerequisite of the prioritizing algorithm. This information is utilized by predictions for re-

serving guard capacity in the literature to track the speed and moving direction of mobiles. However, Wang *et al.* [64] claimed that such information procured from mobility models or GPS monitoring is generally costly and inaccurate, and complicated as well. As an alternative, they proposed using measured pilot strength to predict handoff (in IS-95 systems) since it is the origin of every handoff thus is accurate. Moreover, it is inexpensive since no additional network signaling is needed. We take advantage of this idea for the prediction in our algorithm, but modify it for 3G CDMA systems (*i.e.*, WCDMA). It must be noted that the prediction method introduced in this paper is not as complex and precise as the aforementioned one because our focus is not on guard capacity reservation algorithm. However, with elaborately designed prediction scheme the significance and effectiveness of our algorithm will be more prominent.

Typically, in addition to the avoidance of signaling flood, prediction is updated at the end of every prediction window W_t to remove withdrawals (*i.e.*, (4.5) holds) resulting from incorrect predictions or call termination ($T_c > D_{th}$, see Section 4.4.3 below). The output priority queue (PQ) is updated accordingly based on the latest information procured through prediction notification from mobiles. When handoffs actually take place, mobiles which are in PQ are identified by BS and are allocated channels immediately if the guard power allows. On the other hand, if the handoff requests are not identified as in the regular handoff procedure, these requests have to be sent to the target cell first since the BS has to inform the target to reserve power resources, where there exists the uncertainty about whether these requests can be approved with sufficient resources. Hence with the prediction, the availability of resource is assured to maintain dropping performance. The handoff execution delay is also shortened which may cause power outage and fade margin enlarging [65]. Note that it is wise to shorten this delay by all means especially in our case. Since additional handoff execution time can be caused by queuing and sorting the handoff predictions in the proposed algorithm, which may introduce computation complexity to the base station and reduce the base station's handoff processing speed. All of the above reasons reinforce the need for prediction.

A Predicted Set is proposed in our algorithm, which consists of BSs satisfying the in-

equality beneath,

$$M_{ps}^{b} > \lambda(Best_{ps}^{active} - AS\_Th + AS\_Th\_Hyst). \tag{4.6}$$

The prediction threshold PS_Th obeys the dynamics of the threshold for the active set switching, and is related by $PS\_Th = \lambda(Best_{ps}^{active} - AS\_Th + AS\_Th\_Hyst)$, where λ, $\lambda \in (0,1)$, is a design constant affecting the prediction threshold above which the pilot is added into the predicted set, relative to the active set threshold. (4.6) serves as a trigger for the execution of the prioritizing algorithm. When (4.6) is satisfied, MS will report to BS of the prediction and the call holding time T_c, and the request will be put into the priority queue. As long as the queue is not empty, BS will perform the algorithm at the end of W_t.

4.4.2 Downlink Transmission Power

Next, channel condition should be taken into account of the profile, in that it is the indicator of required handoff power. A user experiencing better link gain and hence demanding less power is given a higher priority, in order to get more users serviced with the same amount of scarce downlink power resource. Assuming the maximum size of the active set is 2 (*i.e.*, at most 2 BSs co-serve a handoff user at the same time), we can apply the maximal ratio combining strategy to (4.1) to derive the E_b/I_0 of a mobile i within the soft handoff zone as:

$$\gamma_i = \sum_{b=0,1} \Gamma \frac{P_{bi}G_{bi}}{(P_T - P_{bi})G_{bi} + \sum_{j \neq b} P_T G_{ji}}, \tag{4.7}$$

where 0 and 1 are in general the two co-serving base station's identity numbers, P_{bi} is the transmission power to mobile i from BS b (current BS 0 and target BS 1). Based on the straightforward power division strategy [72] (*i.e.*, $P_{0i} \approx P_{1i}$), under the presumption of $\sum_{j \neq 0} G_{ji}/G_{0i} \approx \sum_{j \neq 1} G_{ji}/G_{1i}$, the required handoff power from BS1 to mobile i can be written as

$$P_{1i} \approx \frac{\gamma_i P_T (1 + \sum_{j \neq 1} G_{ji}/G_{1i})}{2\Gamma + \gamma_i}. \tag{4.8}$$

4.4.3 Call Holding Time

The last term included in the profile is the call holding time T_c. This information can be easily derived by the UE (user equipment) through monitoring the connection time elapsed for the ongoing call. For the proposed profile, we import a parameter D_{th} denoting the death threshold for ongoing calls. The ongoing call is presumed to be terminated by the user before the actual handoff takes place if its T_c is greater than D_{th} at the time the prediction is made. If $T_c < D_{th}$ holds at the time of prediction, higher priority is assigned to a longer T_c. Because it is more probable that this mobile will terminate its call soon and release the resource for other mobiles' use.

We finally conclude the adaptive priority profile for user i as:

$$AP_i = \frac{1}{\mu P_{1i}^{nor}} + T_{ci}^{nor}, \tag{4.9}$$

in which AP_i is user i's priority and μ is the adaptive factor adjusting the proportion of power and time to be comparable quantitatively. P_{1i}^{nor} and T_{ci}^{nor} denote the normalized downlink transmission power and the normalized call holding time of user i, respectively. $P_{1i}^{nor} = P_{1i}/P_{mean}$, where P_{mean} is the mean downlink transmission power of predicted handoff users for the same target cell. $T_{ci}^{nor} = \lceil T_{ci}/T_{scale} \rceil$, where $T_{scale} = 10s$ is the scale of the calling time. We use a rough calling time measuring method. The available T_{ci}^{nor} values are $0, 1, 2, ..., D_{th}/T_{scale}$. While the actual T_{ci} values can be any number between 0 and D_{th}, for simplicity, we assume that the T_{ci}/T_{scale} values will be ceiled to one of the above T_{ci}^{nor} values for the AP calculation. This definition style is derived from the wired networks, where IGRP and EIGRP routing protocols define a composite metric associated with each route in an alike fashion as mentioned in Chapter 3. Specifically, we subdivide users into two classes, which are distinguished by different priority profiles. According to Viterbi *et al.* [73], the maximum fade margin (Max. γ_d) put apart for overcoming shadowing correlation (with coefficient a^2) is obtained at the cell boundary, subject to a certain outage probability target (P_{out}^*). Hence, we issue boundary users a lower μ since they require a higher power for handoff (due to a higher γ_d) to ensure fairness. For convenience, we set $\mu = 1$ for ordinary

users and $\mu \in (0,1)$ for marginal users. Dedicated surveys on fade margin improvement and delicate relations among parameters such as γ_d, P^*_{out}, and a^2 are present in [65, 67, 73].

4.4.4 The Proposed Handoff Procedure

The implementation procedure of the proposed soft handoff algorithm is drawn in Fig. 4.3. Note that we provide the option of dynamic channel reservation mechanism in the flowchart,

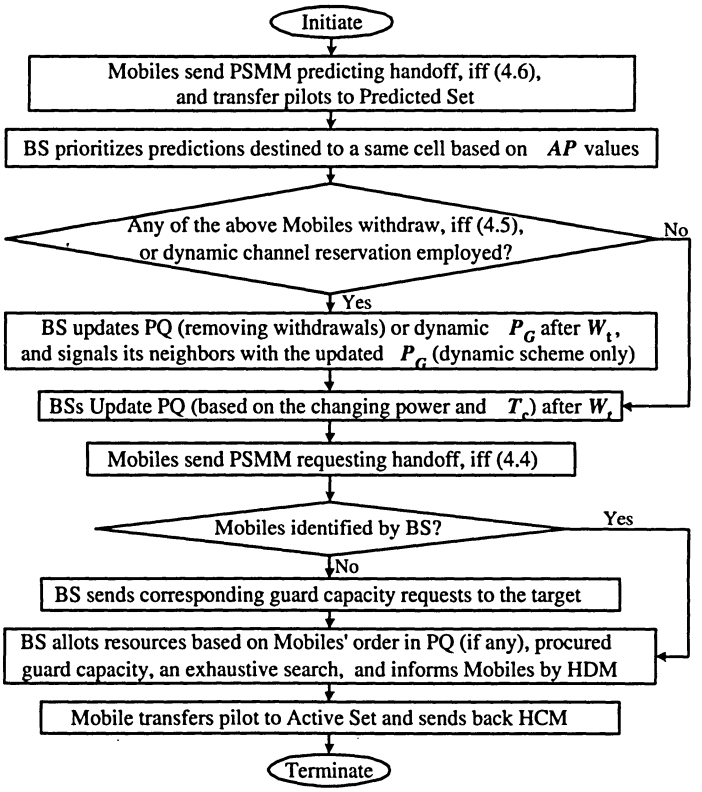

Figure 4.3: Proposed Soft Handoff Procedure.

in spite of its absence in our algorithm. Additionally, the exhaustive search allocation scheme incorporated in the flowchart can be traced in [61], where we proposed a modified queuing algorithm, considering that a user with a smaller required power is possible to be at the back of PQ, since the synthetic AP value is determinant when prioritizing incoming users. While in the classic first-in-first-out queuing scheme, users behind will not be allocated until all of

the front users are serviced.

4.5 Simulation Senario And Results

Table 4.1 displays the system and handoff parameters for our simulation senario. Handoff parameters such as $W_t/\lambda/P_G$, μ (derived from normalized R_s, the radius of non-soft handoff zone in [66]), and $a^2/\gamma_d/P^*_{out}$ are obtained from [64], [66], and [73], respectively. In [64], a fixed guard capacity reservation scheme is shown to have similar performance to their dynamic ones in terms of new call blocking probability, handoff dropping probability, and average power utilization but with higher guard power waste at lighter load. Therefore, we employ fixed reservation scheme for simplicity and generate heavier traffic load (50%-100%) to diminish the guard power utilization discrepancy. Offered traffic load used in the following figures is the actual amount of traffic normalized by the fully loaded amount of traffic in the network. Heavier traffic load imposes greater challenge on the success of the proposed algorithm because of more severe dropping condition.

Table 4.1: Simulation Parameters of The Proposed Handoff Algoirthm

SYSTEM PARS.	VALUE		HANDOFF PARS.	VALUE
Bandwidth/ChipRate	5MHz/3.84Mcps		AS_Th/AS_Hyst/ΔT	2dB/2dB/0
Cell Radius	1000m (macro-cell)		W_t, D_{th}	1s, 2min.
User Locations/Arrival Type	Uniformly/Exp. Distributed		T_c (Exp. Distributed)	mean 1min.
P_T, P_t	20W, 70%P_T		P_G (for each neighbor)	4%P_t
Traffic Model	ON-OFF, 50% ON Prob.		λ, μ	0.85, 0.8
User Mobility/Max. Speed	2-dimensional Random Walk/100Km/h		a^2	0.5
R_v Set	{64, 32, 16, 8}$kbps$		Max. γ_d	6.2dB
$\alpha/\sigma_X/\gamma^*$	4/8dB/5dB		P^*_{out}	0.1

Figs. 4.4-4.6 demonstrate that the proposed prioritizing algorithm outperforms the FG (fixed guard capacity) scheme [64] in terms of handoff dropping probability (P_d), average guard power efficiency, and average guard power utilization, under the same prediction scheme and general handoff procedure proposed in Section 4.4. These performance measures are defined below, where N_{tho}, N_d, N_{ssho}, P_{rsvgrd}, and P_{sumho} denote total number of

received handoff calls, the number of blocked (dropped) handoff calls, the number of successful handoff users, total reserved guard power, and total consumed guard power (sum of successful handoff powers derived from (4.8)), respectively.

- P_d-N_d/N_{tho}.
- The guard power efficiency-N_{ssho}/P_{rsvgrd}. It indicates the number of successful handoff users that can be supported by consuming certain P_{rsvgrd}, thus clarifies the efficiency of P_{rsvgrd}.
- The guard power utilization-P_{sumho}/P_{rsvgrd}. It indicates the utility of the reserved guard power. If it is too low, that implies the underutilization and a waste of system resources.

We run over 10000 trials and get the averages of the above performance indicators as seen in Figs. 4.4, 4.5 and 4.6. The reason for the outperformance has two folds which reflect the innovation of the proposed prioritizing algorithm: (1) handoff users demanding smaller powers are scheduled first in general, contributing to a larger number of successful handoff users with the same amount of guard power, thus P_d is reduced (Fig. 4.4) and power efficiency is enhanced (Fig. 4.5).

(2) Handoff users who have been connected for a longer call time ($T_c < D_{th}$) are more likely to cease and release resources shortly, before other concurrent handoffs are dropped due to the lack of enough guard resources. It can be interpreted as the resource borrowing (or reuse) mechanism of earlier and faster handoffs from slower handoffs, giving rise to higher guard power utilization (Fig. 4.6).

Note that an over-low λ triggers more frequent and false predictions, because it extends the active set and may lead to too weak potential BSs. While an over-high one hinders potential true predictions. Consequently, this parameter has to be designed carefully. The authors [64] addressed similar concerns on λ. We use $\lambda = 0.85$ as indicated in [64].

From Figs. 4.5 and 4.6 we observe that higher offered load yields larger difference between the proposed algorithm and the FG. Indeed, whether the guard capacity reservation scheme

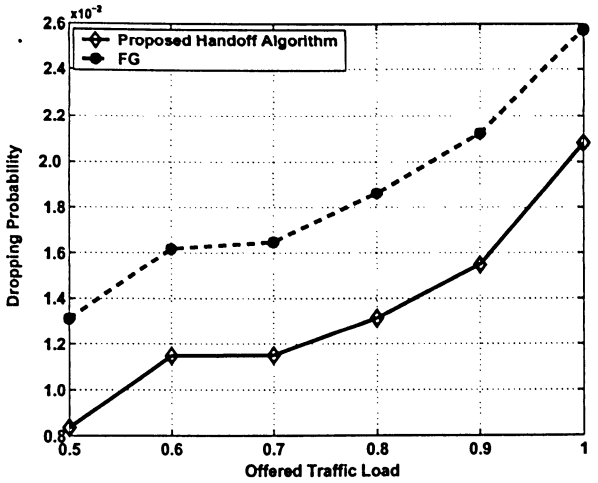

Figure 4.4: Handoff Dropping Probability.

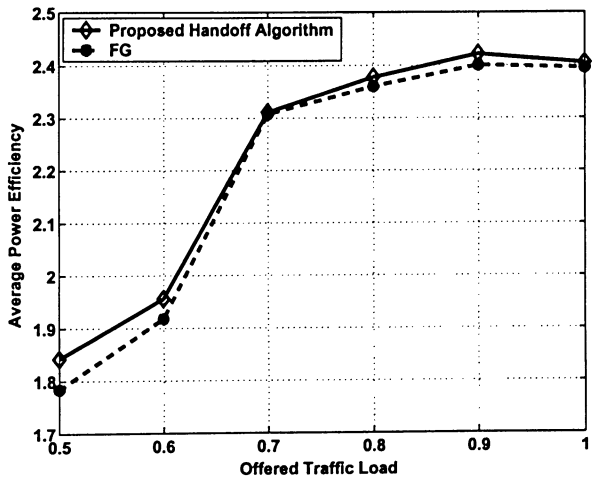

Figure 4.5: Average Power Efficiency for Handoff Algorithm.

is fixed or adaptive produces no distinction in terms of average guard power utilization, as shown in [64]. However, the fixed and adaptive reservation schemes can affect the handoff dropping probability to have very different performances, which is verified also in the above research work. Additionally, the authors show that the adaptive reservation schemes exhibit greater difference between each other as the offered traffic load grows. But since we employ

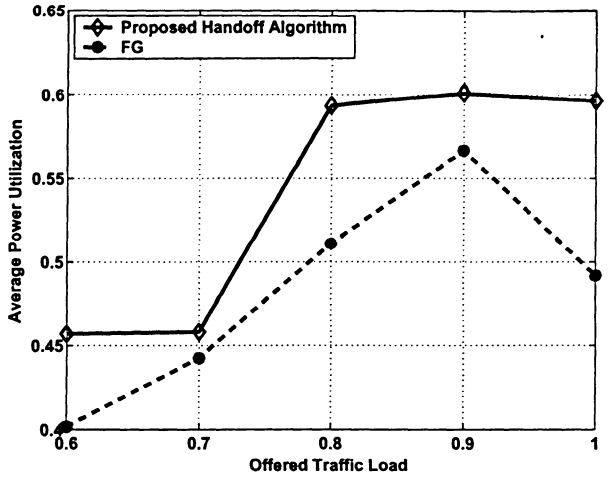

Figure 4.6: Average Power Utilization for Handoff Algorithm.

fixed reservation schemes, the fixed guard power may fail to exploit or adapt to the dynamics of the handoff dropping probability when the system is highly loaded and has poorer dropping behavior. This explains the reason for the nearly parallel curves in Fig. 4.4. In the future research, more performance gains of dropping probability are expected with the employment of the adaptive guard capacity reservation scheme in the proposed handoff algorithm.

Chapter 5

Transport Layer: TCP Performance

The Transmission Control Protocol (TCP) is a connection-oriented, reliable end-to-end transport layer protocol designed for supporting applications and communications in distinct but interconnected networks. The main purposes and fundamental operations of TCP are listed below. TCP establishes connection-oriented communications identified by source and destination IP addresses and port numbers using the well-known three-way handshake mechanism. TCP also renders a reliable connection by Acknowledgment (ACK) packets and flow control using sliding window protocol. Finally, TCP avoids congestion in the network by congestion control which consists of four algorithms slow start, congestion avoidance, fast retransmit, and fast recovery.

The TCP performance concerned in this work is the impact of ACK timeout on the behavior of TCP congestion window (cwnd). When a sender transmits a TCP data segment to a receiver, the sender specifies the sequence number of the next data octet it is expecting in the "Acknowledgment Number" field of the TCP header. At the same time, the sender puts a copy of the transmitted data into a retransmission queue and starts a timer. If it does not receive an ACK for the transmitted data before the timeout, the sender will interpret it as a loss due to network congestion, reduce the congestion window size and retransmit the data. This procedure takes place in the wired part of hybrid wired/wireless networks as well. However, the timeout can be spurious because of some inherent problem of wireless techniques, causing unnecessary TCP throughput degradation. In this chapter, we discuss

this problem, present existing solutions and provide our solutions.

5.1 TCP Performance Degradation Over Wireless Networks and Solutions

5.1.1 TCP Over Wireless Networks

Wireless links will play an important role in future communication networks due to the increasing popularity and the widespread deployment of wireless networks. Communication over wireless links is often characterized by limited bandwidth, high latencies, sporadic high bit-error rates, and intermittent connectivity due to user mobility and handoffs [74, 75]. TCP congestion control is originated and well-investigated in wired networks where congestion is the main cause of packet loss, thus operates properly in such networks. But wireless networks and mobile terminals feature a large amount of losses due to bit errors and handoffs, thus are in some facets non-cooperative with traditional TCP congestion control, resulting in end-to-end performance degradation. For the reliability of the connection mentioned earlier, the TCP sender maintains a running average of the estimated round-trip delay and the mean linear deviation. The sender identifies the loss of a packet by the arrival of several duplicate cumulative acknowledgments. The absence of an expected acknowledgment within a predefined timeout interval is also considered as the loss. The timeout interval T_o is equal to the sum of the smoothed round-trip delay $\overline{RTT}$ and four times its mean deviation D, $T_o = \overline{RTT} + 4D$. In wired networks, TCP assumes that packet loss is caused by congestions and reacts to it by decreasing cwnd, retransmitting the missing packets, triggering congestion control/avoidance mechanism (*i.e.*, slow start [76]), and recalculating the retransmission timer with some backoff according to Karn's algorithm [77]. In wireless networks, when packet loss occurs for some reasons other than congestion, such as temporary blackout due to fading, or when packets are correctly received but the corresponding ACKs have not been returned which is the so-called spurious timeout, TCP will perform the same as for reacting to congestion in wired networks because it is not able to identify these different types of losses. The misbehavior of TCP in wireless communications eventually leads to needless

throughput drop and inefficient bandwidth utilization.

IEEE 802.11 and CDMA are the two most popular wireless standards. In the IEEE 802.11 wireless standard, the MAC layer CSMA/CA (Carrier Sense Multiple Access with Collision Detection) protocol is designed to minimize collisions by using request to send (RTS), clear-to-send (CTS), data and acknowledge (ACK) transmission frames in a sequential fashion. Nevertheless, the "hidden node" problem acts as the major limitation which can disrupt a significant number of communications especially in a highly loaded wireless LAN. Although the TCP handshake mechanism reduces the probability of "hidden node" collisions, it cannot eradicate them. As a result, collisions in the IEEE 802.11 wireless networks become the main reason for TCP performance degradation.

IEEE 802.11 infrastructure differs from CDMA in that all 802.11 compliant products utilize the same PN code and therefore do not have a set of codes available as is required for CDMA operation. Though in CDMA standard the "hidden node" problem is solved, other inherent problems may affect TCP performance, one of which can be induced by the link layer scheduling. Wireless links are known to be time-varying and fading-dominant, challenging the strategy and techniques of scheduling. To confront the problem of wireless channels, various scheduling algorithms have been devised, one of which is the famous wireless opportunistic scheduling aiming to increase the aggregate system capacity by opportunistically scheduling simultaneous users at time instants when the channel qualities are favorable comparatively [78]. A well-known algorithm that achieves precisely this objective is Proportional Fair (PF) which is, for example, used in CDMA 1X EV-DO systems [79]. Our proposed link layer scheduling scheme in this thesis falls into the category of opportunistic scheduling, as does PF. The interaction between TCP and its congestion control mechanism and such scheduling algorithm has been investigated in [78] and [79], both of which show that the scheduling algorithm can introduce significant inter-scheduling intervals, thus stochastic data transfers and highly varying packet transmission times. The variability of the packet transmission time is shown to be large enough to cause spurious TCP timeouts [79], which are particularly outstanding with wireless opportunistic scheduling and unexpectedly initiate

TCP congestion control affecting the performance (*i.e.*, throughput).

5.1.2 Solutions Against TCP Performance Degradation

Although there are difficulties implementing TCP in wireless networks, so far no single researcher has proposed to replace TCP with other transport layer protocols suitable for communications over wireless links. It is unwise to remove TCP since its hierarchical relationship with popular application-layer protocols such as HTTP, FTP, TELNET, and SMTP, has been well established. In order to facilitate the communications between wired and wireless networks, the wireless network applications must use TCP. Therefore, We turn to the option to improve TCP performance over wireless links without modifying the TCP infrastructure, one major benefit being the seamless integration of mobile communications through wireless networks with the wired Internet backbone.

In general, the proposals found in the literature can be categorized into three classes: split-connection protocols, end-to-end protocols, and link-layer protocols. We introduce them individually.

Split-Connection Protocols: Like the well-known Indirect-TCP (I-TCP) [80], protocols falling into this class have a general feature of splitting the end-to-end connection into two connections, with one between the TCP sender and the base station and the other between the base station and the mobile host. The classic TCP remains running in the wired part while the wireless loss recovery is performed independently and is transparent from the TCP sender's viewpoint. This approach brings some desirable results since both networks achieve the optimal performance individually. However, from the whole system's point of view, Balakrishnan *et al.* [74] argue that the approach also introduces drawbacks. For instance, it is likely to have the ACK arriving at the sender even earlier than the receiver's receiving the intended data packet due to the separation of the TCP connections into two independent ones. Furthermore, the application relinking (special socket needed), packet overhead (as packets incur overhead twice at the base station) and handoff latency (a large number of per-TCP-connection information stored at the base station, lowering its processing speed)

indicate the ineffectiveness of this approach.

End-to-End Protocols: This type of protocols are logically in line with the conventional TCP philosophy, with modifications in certain operations and mechanisms such as to recover congestion faster and to be able to detect multiple losses in a single window (*i.e.*, SACK [81]), and to prevent false trigger of congestion control at the sender by using explicit congestion notification (ECN) [82] to enable TCP to recognize the congestion loss from other kinds of losses. An advantage of this approach, contrary to the split-connection approach, is that it sees the entire connection and thus can optimize the overall system performance.

Link-Layer Protocols: This class of protocols handles the wireless link losses at the link layer independently using techniques such as forward error correction (FEC) [83] and local retransmission. They do not rely on transport layer protocols and attempt to solve local problems locally. As a result, the excessive memory of the base station to maintain the per-TCP-connection information is not required. But other problems may arise since these link-layer protocols usually cannot completely block wireless losses from TCP and will threaten TCP performance. Recently, a special protocol, the Snoop protocol, has been proposed [74] and has attracted attention. Its logic is similar and can be categorized to the link-layer protocol class. The difference is that it is not completely isolated from TCP but utilizes some information from TCP layer to overcome the problem of pure link-layer protocols. So the Snoop protocol is actually a TCP-aware link-layer protocol [75]. The advantages and disadvantages of the link-layer protocols also exist in the Snoop protocol since these protocols are the same in nature. But because of its TCP-aware nature, the Snoop protocol improves the TCP performance over wireless links compared to the conventional link-layer protocols.

One may refer to [75] for a detailed survey on different classifications of TCP-over-wireless solutions, where Balakrishnan *et al.* did an extensive and thorough comparison of the above solutions, indicating the benefits and drawbacks. The following paragraph summarizes some of the useful results from their work.

A reliable link-layer protocol that uses knowledge of TCP to shield the sender from duplicate acknowledgments arising from wireless losses gives higher throughput than one link-

layer protocol that operates independently of TCP and does not attempt in-order delivery of packets. Of the schemes investigated, the TCP-aware link-layer protocol with selective acknowledgment performs the best. The split-connection approach, with standard TCP used for the wireless hop, lowers the "wireless" effect on TCP performance but resulted in poor end-to-end throughput due to timeouts on the wireless connection. Furthermore, the authors demonstrate that splitting the end-to-end connection is not a requirement for good performance. The SMART-based selective acknowledgment scheme in a LAN environment and the SACK scheme in the WAN experiments resulted in significantly improved end-to-end performance. These results confirm that selective acknowledgment schemes are very useful in the presence of lossy links, especially when losses occur in bursts. The end-to-end schemes (*i.e.*, ECN) are shown to be promising since significant performance gains can be achieved without any extensive support from intermediate nodes in the network.

5.2 Proposals For TCP Performance Improvement

To the best of our knowledge, the impact of downlink scheduling on the performance degradation of TCP in CDMA networks has not received much research attention. Two works regarding similar issues in time-slotted networks have been found in the existing literature. [78] proposes a reservoir mechanism at the base station to store some ACKs during scheduling mid-season and release them in the off-season to avoid spurious timeouts at TCP sources. It is a revised version or an addition of the Snoop protocol. The problem of this algorithm is that they use ICMP packets to measure the round trip time for ACK release interval calculation. These extra ICMP packets can significantly increase the network traffic especially in a large network where there are lots of TCP senders and receivers. In addition, they did not demonstrate clearly what methodology they utilized to measure the idle period and the scheduling cycle at the base station. [79] proposes to use pure MAC layer information to calculate a TCP-related metric for link layer scheduling. Thus TCP performance is maintained when this metric is used in the link layer to schedule traffic from TCP sources. This algorithm can also be called TCP-aware link-layer algorithm. A crucial part of this

algorithm is to use MAC information to approximately calculate the average RTT. However, this approach is very complicated since it requires heavy mathematical calculations to obtain the new metric at the beginning of each scheduling cycle and to update the information for recursive calculation afterwards.

One of the innovations of our works is that we propose two alternative methods to combat TCP performance degradation due to wireless scheduling in CDMA downlinks. In this research, we do not entail to explain that wireless opportunistic scheduling impacts TCP performance since it is well illustrated in [78] and [79]. We assume it is the existent problem and are concerned with the solutions in a CDMA environment. In TDMA and systems alike (used by [78, 79]), scheduling is necessary because time slot is contended (the user with the best channel condition is selected) before every scheduling cycle. In CDMA systems, however, such scheduling is not required since different users can transmit simultaneously using distinctive codes, which is one of the advantages that make CDMA preferable. Nevertheless, as we addressed in the preceding chapter that in the downlink direction, it holds that scheduling users under better channel condition first can improve overall network performances. Wherever scheduling arises in wireless networks, there are impacts on TCP sources.

The proposals are based on the principles of real-time video transmissions where the jitter is smoothed out at the receiver to ensure a constant rate playout. We apply this idea to CDMA downlink scheduling which introduces stochastic halt affecting TCP performance at the sender (typically a fixed station in the wired part of the network from which data can be downloaded using HTTP or FTP application protocols). Fig. 5.1 depicts the network topology we are dealing with, where RNC (Radio Network Controller) and PDSN (Packet Data Serving Node) connect the wireless network with the Internet backbone.

TCP mechanism is quite mature in wired networks. The problem that we are facing now is because of the exclusion of potential wireless applications when TCP was proposed. Naturally, one solution would be emulating the behavior of the wired network, so that the "wireless" effect on TCP could be eliminated. Wired scheduling is periodical and hence

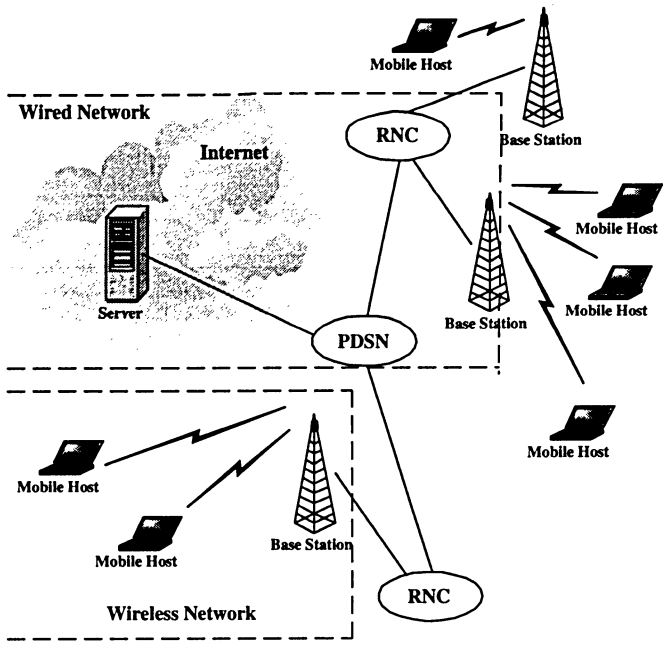

Figure 5.1: The Integrated Network Topology for TCP Proposals.

predictable since every user is equal in terms of channel condition (no time-varying fading over wired links). It can be prevented and will not be a cause of TCP spurious timeouts. Thus wired scheduling is not within the scope of our study. On the other hand, wireless scheduling is unpredictable and thus irregular due to time-varying wireless links (users have to be re-scheduled according to their instant channel fading). Wireless CDMA networks consist of two parts: the uplink and the downlink. In the reverse direction (uplink, *i.e.*, from the mobile station to the base station), the key restriction is the incremental interference in the system as communicating mobiles increase, due to transmission power levels of other active users and imperfect orthogonality of channel codes. Scheduling in this direction is not needed as long as the system interference stays below the threshold. Here we assume that the simultaneously active users in the system are not enough to cause the interference beyond the threshold. Therefore, scheduling in this direction is of little importance to be considered by TCP performance. Rather, we focus on the downlink direction where we proposed novel

scheduling schemes and explained their necessity and effectiveness in Chapter 3.

5.2.1 Proposal 1

Through the analysis above, downlink scheduling is the only affecting factor to degrade TCP performance in our study. Specifically, when inter-scheduling cutoffs (intervals/halts) occur, there is a temporary silent period in the wireless part of the network for the scheduler to collect the up-to-date channel information and to perform the new scheduling at the base station. During this period, no traffic is in the wireless network and the mobile station will not send back the expected ACK since it has not received the TCP packet queued at the base station. Consequently, the TCP source may undergo spurious timeouts without the ACK it is expecting.

What if we avoid this burst-and-silence traffic pattern to smooth the traffic throughout the burst and the following silence period, just as what we do to avoid annoying jitter in video playout? Then the TCP packets stored in the base station will arrive at the mobile station with steady rate and the mobile station will return the ACK without huge gaps for TCP to timeout. After the scheduler determines the order of the packets to be transmitted based on the channel conditions of each mobile user, it calculates a new transmission rate to send the packets in a steady pace instead of sending them out all at once. In this case, there is traffic flowing in the network at all times so that there is no cutoff any more. This is our first proposal, Proposal 1. Let the queues at the base station be per-TCP-flow and the ACK be per-packet based (TCP Reno). Let N_i be the number of TCP packets in queue i of the reference base station, let T_{bi} and T_{si} (in seconds) be the midseason (burst) and offseason (silence) duration of the scheduling cycle of queue i, respectively. The playout rate to smooth out the "jitter" of the burst traffic R_p is written as:

$$R_p = \frac{N_i}{T_{bi} + T_{si}}, \tag{5.1}$$

where N_i is known to the base station through queue monitoring, $(T_{bi} + T_{si})$ is equivalent to one term T_{ci}, the scheduling cycle of queue i, which can be obtained from the history as:

$$T_{ci}(n) = (1 - \rho)\overline{T}_{ci}(n-1) + \rho T_{ci}(n-1), \tag{5.2}$$

where $T_{ci}(n)$ and $T_{ci}(n-1)$ denote the nth and its previous, the $(n-1)th$ scheduling cycle, respectively. $\overline{T}_{ci}(n-1)$ denotes the average duration of scheduling cycles of queue i up to scheduling cycle $(n-1)$. ρ is a weighing parameter with a typical value of $\frac{1}{1000}$ [79]. The initial value of the scheduling cycle (*i.e.*, $T_{ci}(1)$) can be monitored by the base station through some timer setting. Having smoothed out the "jitter" using the above algorithm, the base station can "play" the traffic continuously and gets the ACK back to the TCP sender accordingly, without temporary blackout which is the root of TCP spurious timeout and performance degradation.

5.2.2 Proposal 2

The above proposal can be easily implemented and effective, which is based on the fact that the ACK flow back to the TCP source is continuous as long as the TCP packets waiting at the base station get transmitted to the mobile destination continuously. It applies to wireless part of the network with both comparable and negligible delay compared with the delay in wired part of the network. Because the timeout interval is updated by TCP through the measured variable round trip delay. Next, we present an idea to design a new protocol (Proposal 2). This method is relatively complicated because it deploys new standards such as signaling and compatibility of this new protocol within existing protocol stack, instead of an algorithm to obtain necessary information and achieve the desired performance. While the advantage of such new protocol is that the potential effects such as additional packets for inquiring RTT [78] and the complexity of the TCP metric calculations [79] are precluded.

Since it is difficult to propose a new standard protocol based on our limited level, we try to briefly introduce the idea for the new protocol which is motivated by an existing standard protocol. In video conference, for instance, attending senders and receivers run Real-time Transport Protocol (RTP) between the application layer (*e.g.*, NetMeeting) and the transport layer (UDP). For real-time traffic, it is of great importance to understand the effect of delay and jitter to get better audio or video quality while communicating. RTP is an Internet protocol standard that specifies a way for programs to manage the real-

time transmission of multimedia data over either unicast or multicast network services. It was originally designed by the Audio-Video Transport Working Group of the IETF (Internet Engineering Task Force) to support video conferences with multiple, geographically dispersed participants. RTP is commonly used in Internet telephony applications and is independent of the underlying transport and network layers. The main components of a RTP header we are concerned here include: "Sequence Number", which is used to detect lost packets and "Timestamp", which is used to detect different delay jitter within a single stream and compensate for it, for the purpose of intra-media synchronization, which is a noticable benefit of RTP. Consequently, the receiver is able to calculate the playout time and to remove the jitter for that stream to play out smoothly.

Our second proposal is enlightened by the above RTP protocol and its mechanism. However, if it is applied to non real-time data traffic, the protocol should be modified accordingly to fit data characteristics. As explained before, we attempt to get the ACK flow back to the TCP source constantly to prevent spurious timeouts. A new protocol named non real-time transmission protocol (NRTP) can be designed based on the special requirement of ACK flows. Assume that the base station can intercept the ACKs from the mobile station to the TCP source like the famous Snoop protocol [74]. The NRTP can be implemented at the base station to prepend the NRTP header to the ACK data packet. Upon receiving the ACK packets with the information needed in the NRTP header to calculate the proper playout time, the TCP source first puts the burst ACKs in a buffer and then plays it out to TCP according to the playout time. Note that the TCP source tries to keep the ACKs flow through both the burst and the silence period of the wireless scheduling cycle to prevent the spurious timeout at its transport layer. The modified ACK data packet at the base station are shown in Fig. 5.2. The NRTP header format may be similar to the RTP header format. The "Sequence Number" field is used to detect the loss of the NRTP packets. The "Timestamp" field is used for clock synchronization and jitter calculations. Some standard protocols such as NTP (Network Time Protocol) can be employed to obtain the time stamp.

The playout time algorithm can also be similar to the calculations used by RTP. As

MAC HEADER	IP HEADER	TCP HEADER	NRTP HEADER	DATA (ACK)

Figure 5.2: The Modified ACK Data Packet Format.

shown in Fig. 5.3, during a single scheduling cycle, the sender (the base station) transmits the ACK packets back to the receiver (the TCP sender) uniformly.

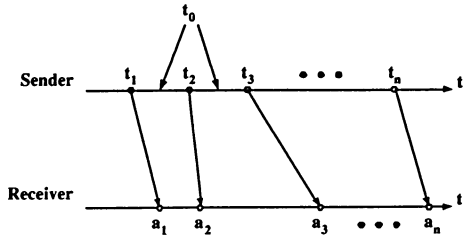

Figure 5.3: ACK Packets at The TCP Sender with Jitter.

The difference of delay times (the jitter) emerges as the packet traverses the network, resulting in non-uniform arrivals at the receiver. In this figure, t_n and a_n represent the transmission and arrival time instants of packet n, $n = 1, 2, 3, ...$, respectively. t_n can be retrieved from the packet header at the receiver (the TCP sender) for further calculation. Upon receiving the ACK packets within a wireless scheduling period (burst and silence), the TCP sender stamps each packet with a time stamp based on the arrival time a, puts them in a buffer and performs the playout algorithm. Let p_n be the playout completion time instant of packet n. Let D_p, t_0, τ and J_{max} be the playout delay, the transmission interval at the sender (the base station), the minimum delay, and the maximum jitter, respectively. The algorithm is derived as follows for each individual scheduling cycle:

1. The following inequality should be satisfied due to the fact that by the time packet n is played out, at least the next one packet should have arrived.

$$p_n \geq a_{n+1}, \tag{5.3}$$

where,

$$p_n = a_n + D_p + t_0, \tag{5.4}$$

2. Also we have the following inequality true:

$$a_{n+1} \leq t_n + t_0 + \tau + J_{max}, \tag{5.5}$$

where $t_{n+1} = t_n + t_0$, and $\tau + J_{max}$ is the maximum delay.

3. In order to satisfy (5.3), the following inequality holds jointly considering (5.3), (5.4) and (5.5).

$$a_n + D_p \geq t_n + \tau + J_{max}, \tag{5.6}$$

where $a_n \geq t_n + \tau$ is obviously the fact.

4. As a result, the minimum requirement for (5.6) to be satisfied for all cases is

$$D_p = J_{max}. \tag{5.7}$$

If the time stamps are known, the maximum jitter of the delays can be calculated, thus the playout delay can be obtained by (5.7) to decide the actual "playout" time for each packet. Note that in this proposal the receiver spreads the ACK pakcets received from the burst period over the entire scheduling cycle to fill in the silent-period blank with ACKs as well. This is the key solution to eliminate spurious ACK timeouts and to enhance the TCP throughput performance consequently.

5.3 Simulations and Performance Improvement

By far we have demonstrated two alternative proposals for TCP to overcome the problem caused by wireless scheduling. This section is dedicated to the numerical results for the feasibility of the proposed algorithm (Proposal 1) for TCP performance improvement, in terms of the evolution of TCP congestion window (cwnd) and TCP throughput . With the reasonable analysis and implementation details, success and effectiveness of the proposed algorithms are further confirmed in the simulation. While the Network Simulator-Version 2 (*ns*-2) [84] models do not support CDMA air interface or CDMA MAC implementations, there is difficulty to simulate the TCP performance with the presence of the proposed scheduling schemes

in *ns*-2. In this thesis, we provide the experimental results using MATLAB and leave the simulations in *ns*-2 for future work. We set up a simple network as shown in Fig. 5.4, where there are a FTP source (node 0) sending TCP (data) traffic and a CBR source (node 1) sending UDP (voice/video, referred as voice hereafter for simplicity) traffic. We create 5 nodes (nodes 3-7) at the wireless terminal as the receivers of both TCP and UDP packets. The base station is modeled by node 2 connecting the bottleneck link. Note that the simulation for Proposal 1 does not include the UDP sender thus studies the TCP performance based on a pure TCP-traffic network. However, the simulation for the extended analysis of the design parameters employs exactly the topology as Fig. 5.4.

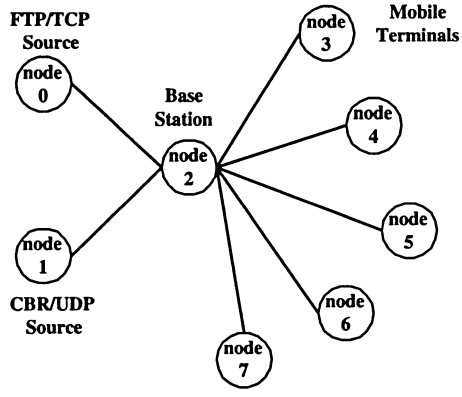

Figure 5.4: Network Topology for TCP-over-wireless Simulations.

The MSs act as the TCP sinks. Assume that the TCP sender always has data to transmit and can transmit as many packets as its transmission window allows (bulk TCP data). The TCP sinks receive TCP packets to deliver them to the user and generate immediate ACKs for the TCP sender. The TCP mechanism implemented for the experiment is only related to dynamic congestion window evolution, according to the slow start and congestion avoidance algorithms [76]. Other mechanisms such as retransmission and recovery for error control are not considered.

The data is transmitted using variable bit rate in the TCP packets with a length of 1000 *bytes/packet*. Voice (only simulated in Section 5.3.2 for design parameters analysis)

is transmitted in the UDP packet with constant bit rate selected from one of the following values: 8 *kbps*, 16 *kbps*, 32 *kbps* and 64 *kbps*, as used in the scheduling schemes in Chapter 3. The packet length of a UDP packet is a design parameter in our research and will be demonstrated later. The buffers proposed in the base station are per-TCP-connection and drop-tail, with the size varied in the simulation. The bandwidth and propagation delay of the wired connections (node 0-node2, node 1-node 2) are 20 *Mbps* and 20 *msec*, while they are set to 1 *Mbps* and 1 *msec* for the wireless connections (node 2-node i, $i = 3, 4, 5, 6, 7$.).

5.3.1 Performance Evaluation of the Proposals

The congestion window behavior of Proposal 1 is compared to that of the standard TCP, TCP Reno with standard ACK (one ACK per TCP packet). The design parameter chosen for this purpose is the buffer size/queue limit BU. We set it to 40 (packets) based on the analysis in the next section 5.3.2. Figs. 5.5 and 5.6 illustrate the evolution of cwnd in terms of the nominal simulation time, for Standard TCP and Proposal 1, respectively. Note that the nominal simulation time is used as the x axis instead of the real simulation time (s) because of the lack of the simulator. We run the simulation created by algorithms in MATLAB for over 60000 times, approximately every 10000 times represents a $100s$ in a real simulator (*i.e.*, *ns*-2). The curves in Figs. 5.5 and 5.6 are plotted using the data obtained from every nominal time ($0.01s$ in *ns*-2).

In order to distinguish the spurious timeout and the real-loss-triggered timeout, we set the wireless link loss (due to shadowing, fading) to be 0. The irregular fluctuation of the cwnd in Fig. 5.5 suggests only the spurious timeout due to wireless scheduling intervals and implies the incompetence of the standard TCP in the presence of wireless scheduling, which is implemented based on the MAPQ and the UF schemes proposed in Chapter 3. We also observe from Fig. 5.6 that compared to Standard TCP, Proposal 1 performs much better in the cwnd evolution because it avoids the spurious timeouts which largely degrade the cwnd performance. We acquire the desired cwnd behavior from Fig. 5.6 that the congestion window keeps increasing until the buffer overflows. Thus, the maximum window size is

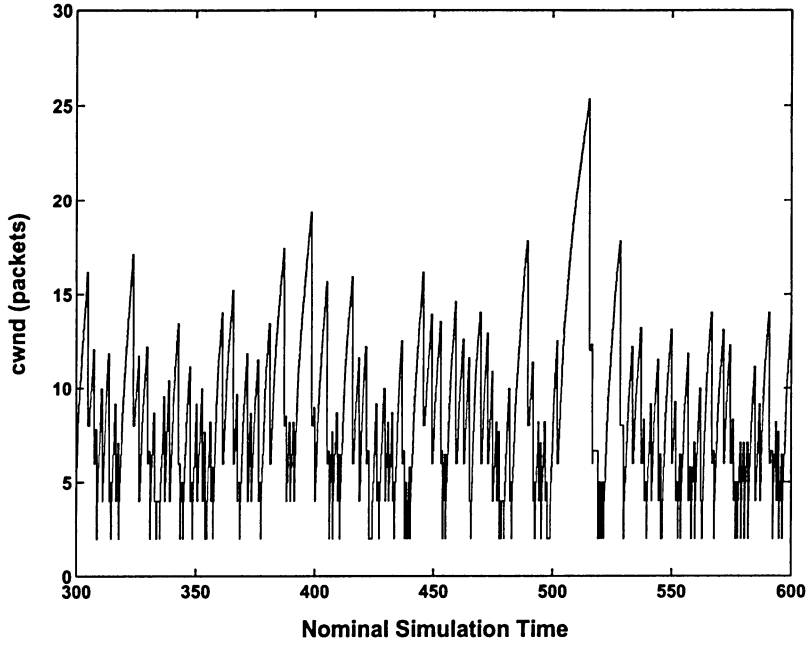

Figure 5.5: Evolution of TCP cwnd for Standard TCP.

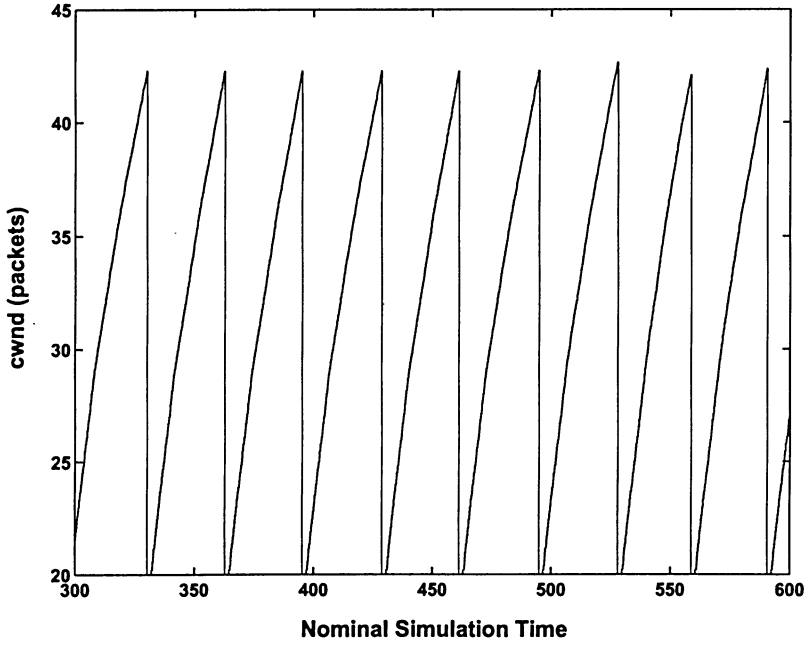

Figure 5.6: Evolution of TCP cwnd for Proposal 1.

Table 5.1: TCP Throughput Comparison

TCP Scheme	Throughput (*kbps*)	# of TCP Packets Generated
Standard	54.216	89149
Proposal 1	70.217	90125

always maintained which is approximate to the buffer size (in our simulation 40 packets). While the congestion window behavior in Fig. 5.5 does not obey the saw-tooth shape, and the achievable window size is even less than 20.

In addition, we analyze the TCP throughput performance for the standard TCP and Proposal 1, as displayed in Table 5.1. The throughput is calculated by

$$Thru_{TCP} = L_{TCP} N_{TCP} / T_{nom}, \tag{5.8}$$

where $Thru_{TCP}$, L_{TCP}, and N_{TCP} denote the TCP throughput (*bps*), the TCP packet length (8000 *bits/packet*), and the total number of TCP packets generated (shown in Table 5.1), respectively. T_{nom} denotes the total nominal simulation time.

We may not see striking throughput gain from this example since the network we built is relatively simple thus not enough traffic is generated to test the throughput behavior. But from the tendency one can clearly tell that Proposal 1 has more desirable throughput performance than the standard TCP. We believe that with heavy loaded network simulated in the future, this throughput gain will be more apparent.

5.3.2 Extended Analysis of the Design Parameters

Although the impacts of the network layer QoS profile, which manages the queuing disciplines and the bandwidth allocation, are beyond the scope of this research, we analyze the basic role in the evolution of TCP congestion window since the queuing mechanism for Proposal 1 at the base station makes use of the network layer functions. We found through the simulation by using *ns*-2 that in general, the variations of PS (voice packet size), $CBRR$ (voice transmission rate) and BU have great impacts on TCP cwnd. In the subsequent simulations, we employ Standard TCP over the wired network (topology shown in Fig. 5.4) with 0 loss rate to study the effects of the design parameters. Figs. 5.7-5.9 depict the change

of cwnd as a result of the varying PS (in the simulation 80, 200 and 1000 *bytes*), with $CBRR$ fixed at 64 *kbps* and BU 30 packets.

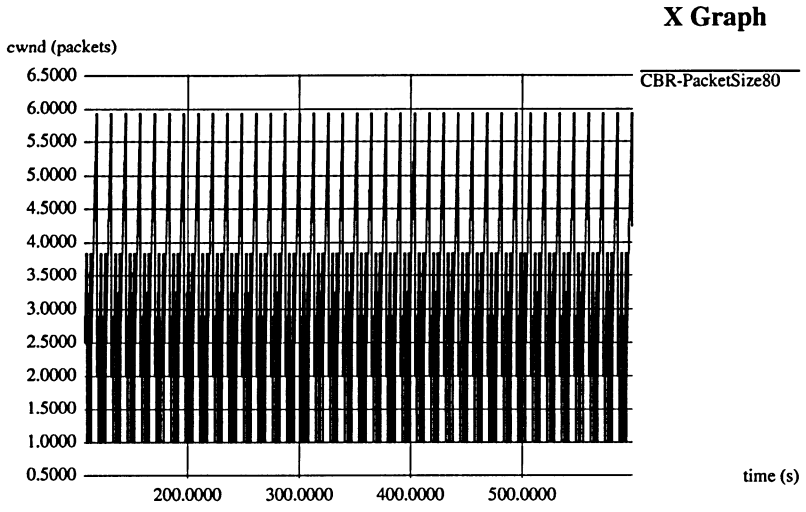

Figure 5.7: Evolution of TCP cwnd with PS=80.

It is noticed that a smaller voice packet size (Fig. 5.7) results in greater reduction of congestion window size, which will largely affect the improvement of TCP throughput. The reason may be that since the wired links in the network have the same bandwidth, the smaller voice packets produced by the CBR source arrive faster and have a higher generation rate. They are of higher probability to be queued in front of the bigger data packets thus get transmitted first. On the other hand, data packets are more likely to be dropped if the queue overflows. This loss of TCP packets will further lead to the loss of the corresponding ACK. It will eventually triggers the TCP timeout and then the congestion window reduction. Note that we did not configure queuing-related parameters (*e.g.*, queuing disciplines) since they are not considered in this research. By default the queue is FIFO and drop-tail.

The effect of $CBRR$ and BU is analyzed in Figs. 5.10-5.13 and Figs. 5.14-5.16. Figs. 5.10-5.13 show the different effects of $CBRR$ when assigned 8, 16, 32, 64, respectively. While PS and BU are assigned 80 and 30, individually. These figures present the rationale explicitly: the higher the voice transmission rate, the more likely that there is congestion

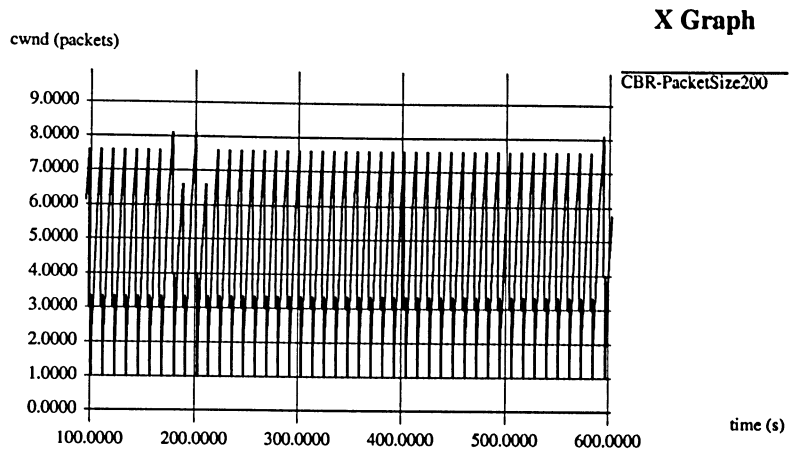

Figure 5.8: Evolution of TCP cwnd with PS=200.

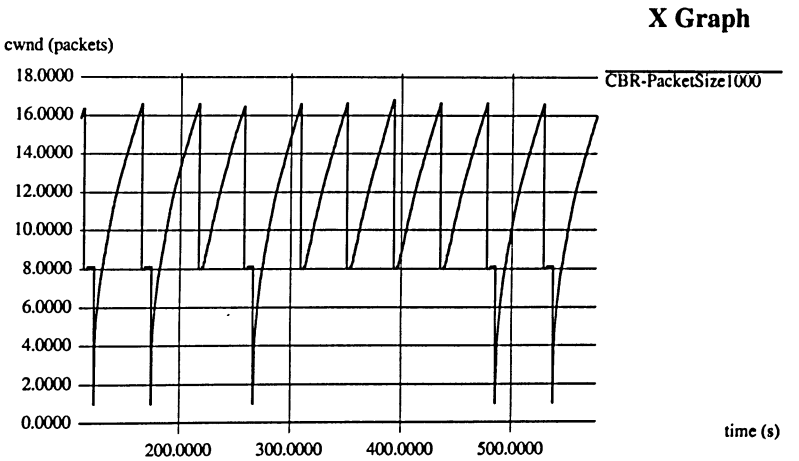

Figure 5.9: Evolution of TCP cwnd with PS=1000.

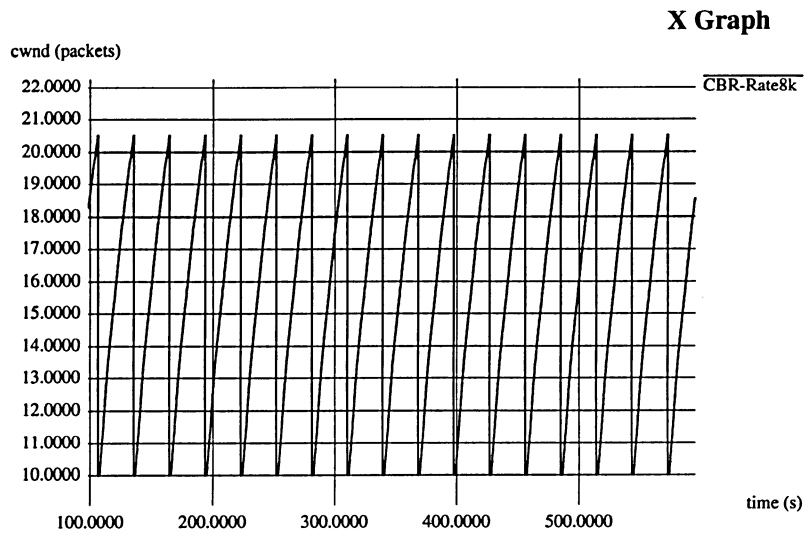

Figure 5.10: Evolution of TCP cwnd with $CBRR$=8.

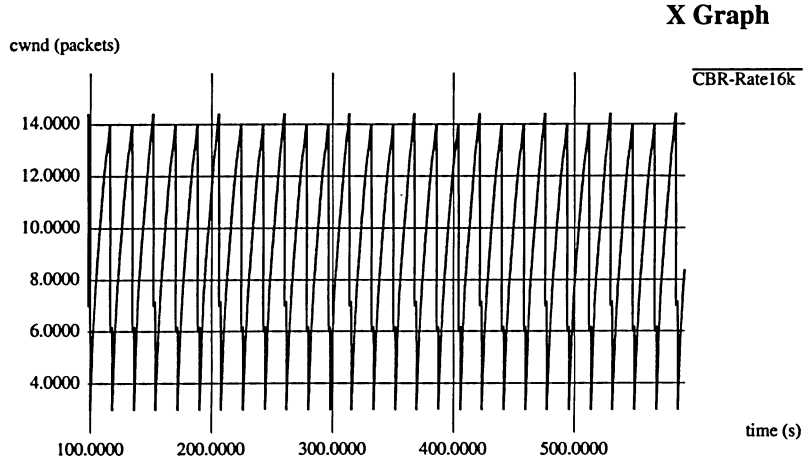

Figure 5.11: Evolution of TCP cwnd with $CBRR$=16.

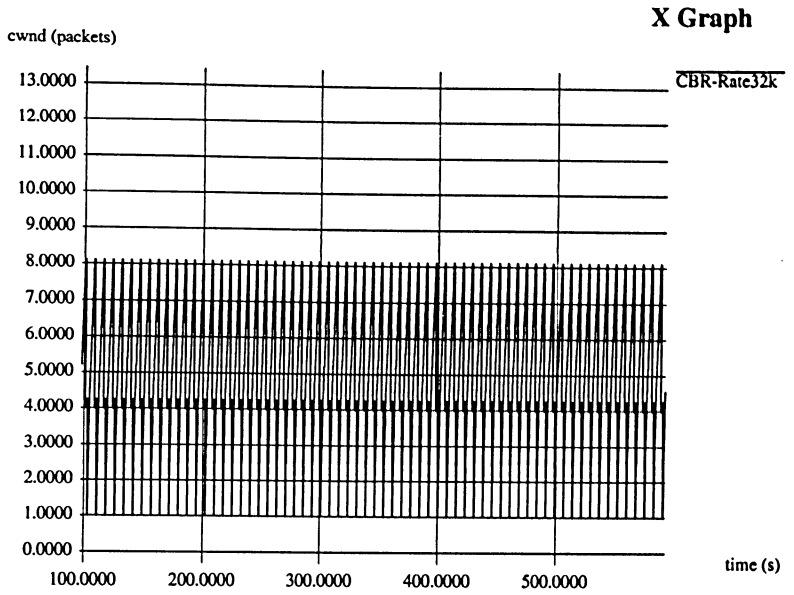

Figure 5.12: Evolution of TCP cwnd with $CBRR$=32.

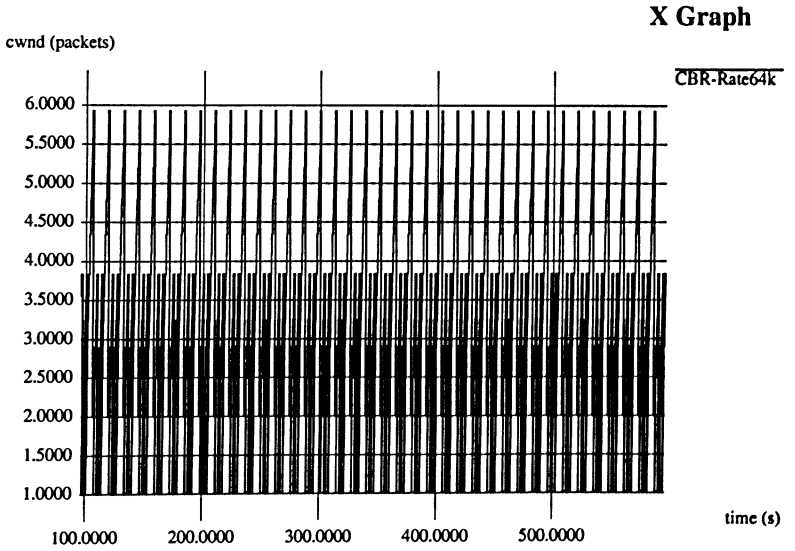

Figure 5.13: Evolution of TCP cwnd with $CBRR$=64.

in the network since the bandwidth resource becomes more demanding. The increasing contention for the available bandwidth will induce more packet loss for both traffic, and the timeouts will occur inevitably which decreases the congestion window size.

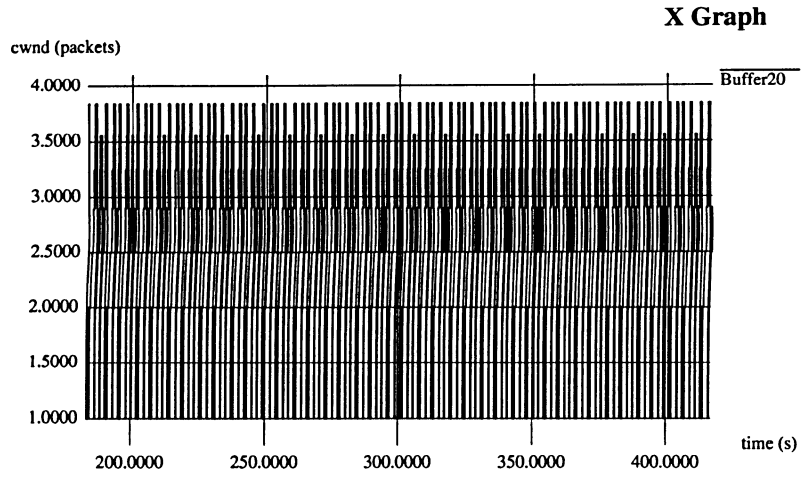

Figure 5.14: Evolution of TCP cwnd with BU=20.

Figs. 5.14-5.16 address the cwnd performance degradation in terms of different buffer size BU, which is assigned 20, 200, and 400. PS and $CBRR$ set here are 80 and 64, respectively. It makes sense from these figures that if the buffer is designed to be small (Fig. 5.14), the memory required for buffering is low trading off the maximum attainable cwnd because the buffer will be full rapidly. In other words, optimizing the congestion window with a large buffer (Fig. 5.16) challenges the excessive memory which can be costly. The buffer selection is hence important as well as complicated and should be designed carefully. Furthermore, we realize through the comparison of Fig. 5.11 and Fig. 5.15 that a lower transmission rate (16 *kbps* vs. 64 *kbps*) requires a considerably smaller buffer (30 vs. 200), yielding similar or even better cwnd performance. Analysis of the buffer size is the only issue involved as the network-layer mechanism that we address in this work. Taking the comparison and analysis of this subsection into account, we designed the simulation parameter BU in the previous section 5.3.1.

From the extended analysis of the design parameters, we further conclude that for an

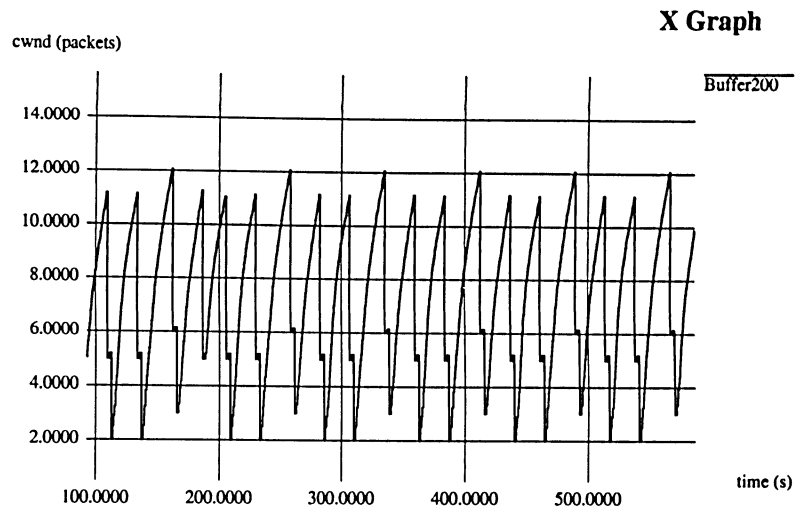

Figure 5.15: Evolution of TCP cwnd with BU=200.

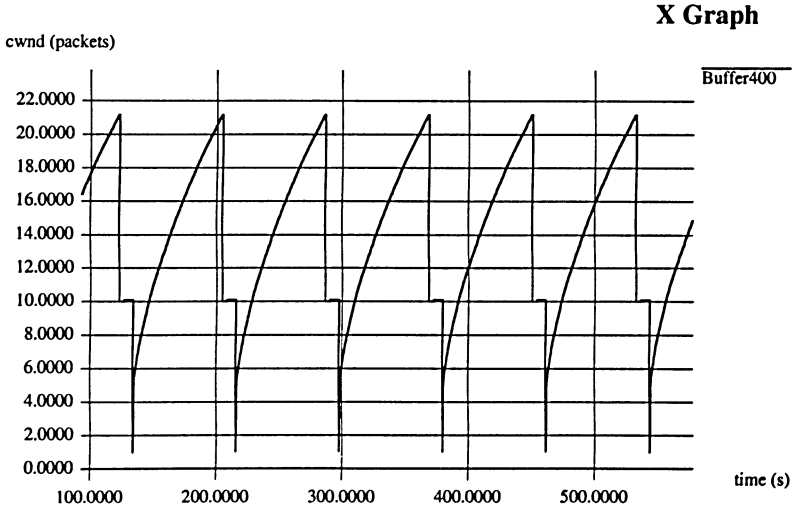

Figure 5.16: Evolution of TCP cwnd with BU=400.

algorithm or a protocol to work appropriately, several key parameters (*i.e.*, PS, $CBRR$, BU) need to be tuned carefully.

The experiment and simulation for this chapter are primarily operated for verifying the feasibility and effectiveness of the proposed strategies. It is preparatory for our future research which includes: in-depth study of the impacts of varying TCP/UDP traffic ratio on TCP behaviors in the integrated wired/wireless environment which were somehow demonstrated in the above simulations; a more realistic network structure and traffic generation for overcoming practical problems, such as the existent "wireless" effect on TCP performance, using *ns*-2.

Chapter 6

Conclusions And Open Issues

We addressed the performance measures and numerical results in the previous three sections, where we proposed strategies and discussed the performance gain for each strategy. We do not attempt to mention these details again, but would like to interpret the gain of our cross-layer design, as the title suggests, to avoid the loss of the topic of this thesis. As argued in Chapter 1, different interpretation of "cross layer" yields different concern on complex connections among layers. Some layers may interact in terms of one measure while others may be related in terms of another. In general, it is difficult to generate a method that conforms the performance measures across all the four layers involved in our research. As proposed in Chapters 3-5, the interacting layers are LINK/PHY, NETWORK/LINK/PHY, and TRANSPORT/NETWORK/LINK, respectively, where they were associated based on currently prevalent and practical problems concerned. Thus for each combination which was formulated by these layers' featured relationships/interactions, there were individual measures that best exhibited the performance gain over non-combination. This is how we designed the simulation senarios to exploit the performance gain for each cross-layer combination.

6.1 Conclusions

In this work, we analyzed and designed cross layer algorithms/schemes to improve overall performance across the entire cellular CDMA network, specialized in downlinks. We pro-

posed a link layer scheduling scheme MAPQ as a cross-layer resource management issue for efficient resource allocation of underlying layer. Evaluation of the proposed scheme has been performed with a reference scheme and superiorities are verified. It should be noted that this scheme can also be used for data scheduling with slight modification of *wdp* (weight of delay over power) and QoS requirement of data traffic.

Considering queuing delay as an important event trigger and service indicator, together with required transmission power/rate, we also proposed an adaptive priority based scheduling algorithm for unified voice/data frameworking and succeeded in fulfilling preset expectation through simulation.

By jointly considering downlink transmission power and call holding time, we proposed an adaptive prioritizing algorithm to control concurrent handoff events to the same destination. The performance improvement was obtained in the simulation.

At the transport layer, we studied its interaction with wireless link layer, particularly, wireless link scheduling. We proposed two methods to avoid TCP spurious timeouts, to regulate the behavior of TCP congestion window, and to enhance the TCP throughput. These methods are claimed to be reasonable by theoretical analysis as well as the simulation verification. Extension of the simulation network and the traffic load scale to obtain greater performance gain of the proposed strategies is left to our future work.

6.2 Open Issues and Limitations

Although the results obtained are what we have expected and are encouraging, there are some open issues and limitations of this work which call for deeper investigation.

1. The performance of the unified voice/data scheduling framework was studied in a 19-cell layout. However, we can further consider the effects of the load variations in the outer cells on the performance in a target cell.

2. In Chapter 3, further details pertaining to the choice of the parameters (a, b and c) will be provided.

3. In the handoff proposal, the proposed algorithm will be more mature if we consider employing recursive looping in the algorithm.

4. Another aspect which is not covered in this algorithm is the additional handoff execution time caused by the introduction of the priority queuing. This can be important when signals from cells in the active set are dropping quite fast.

5. We are seeking for testing the applicability of this algorithm in real systems, regarding the potential signaling overhead.

6. To import dynamic or adaptive guard capacity reservation scheme to the proposed handoff prioritizing algorithm would be an interesting topic.

7. More results will be shown on the role of λ (the design parameter that alters the prediction threshold in the proposed handoff algorithm) and the capacity of base stations.

8. For the TCP proposals, we will build up a larger scaled network where more subscribers and application sources will be generated for deeper understanding of the TCP behaviors. Eventually more efficient designs will be proposed for minimizing the "wireless" effect in wireless networks.

9. In-depth study of the impacts of varying TCP/UDP traffic ratio on TCP behaviors in the integrated wired/wireless environment, a more realistic network structure and traffic generation model for overcoming practical TCP-over-wireless problems.

Appendix A

Acronyms and Abbreviations

CDMA	Code Division Multiple Access
TCP	Transmission Control Protocol
MAC	Medium Access Control
TDMA	Time Division Multiple Access
FDMA	Frequency Division Multiple Access
OSI	Open System Interconnection
WCDMA	Wideband Code Division Multiple Access
PHY	Physical
QoS	Quality of Service
SIR	Signal to Interference Ratio
RTT	Round Trip Time
MAPQ	Modified Adaptive Priority Queuing
UF	Unified Framework
PQ	Priority Queuing
UTRA	Universal Terrestrial Radio Access
3GPP	3rd Generation Partnership Project
DS-CDMA	Direct Sequence Code Division Multiple Access
MS	Mobile Station
BS	Base Station
BER	Bit Error Rate
CLPC	Closed Loop Power Control
OLPC	Open Loop Power Control
VSG	Variable Spreading Gain
IGRP	Interior Gateway Routing Protocol
EIGRP	Enhanced Interior Gateway Routing Protocol
AP	Adaptive Priority
wdp	Weight of Delay over Power
FIFO	First In First Out
pwt	Power Budget

STPD	Scheduling with Transmission Power and Delay
SPS	Static Priority Scheduling
RF	Rate Fairness
GSM	Global System for Mobile Communications
CAC	Connection Admission Control
PSMM	Pilot Strength Measurement Message
HDM	Handoff Direction Message
HCM	Handoff Completion Message
GPS	Global Positioning System
UE	User Equipment
FG	Fixed Guard Capacity
ACK	Acknowledgment
cwnd	Congestion Window
IEEE	Institute of Electrical and Electronics Engineers
CSMA/CA	Carrier Sense Multiple Access with Collision Avoidance
RTS	Request to Send
CTS	Clear to Send
LAN	Local Area Network
PN	Pseudo Random Noise
PF	Proportional Fair
EV-DO	Evolution Data Only
HTTP	Hypertext Transfer Protocol
FTP	File Transfer Protocol
SMTP	Simple Mail Transfer Protocol
I-TCP	Indirect Transmission Control Protocol
SACK	Selective Acknowledgment
ECN	Explicit Congestion Notification
FEC	Forward Error Correction
WAN	Wide Area Network
ICMP	Internet Control Message Protocol
RNC	Radio Network Controller
PDSN	Packet Data Serving Node
RTP	Real-time Transport Protocol
UDP	User Datagram Protocol
IETF	Internet Engineering Task Force
NRTP	Non Realtime Transport Protocol
NTP	Network Time Protocol
CBR	Constant Bit Rate
PS	Packet Size
$CBRR$	Voice Transmission Rate
BU	Buffer Size

Appendix B

My Publications

1. Jin Yuan Sun, Yifan Peng, and Lian Zhao, "A Novel Packet Scheduling Scheme Based On Adaptive Power/Delay for Efficient Resource Allocation in Downlink CDMA Systems," Canadian Conference on Electrical and Computer Engineering 2005, May 2005.

2. Jin Yuan Sun, Lian Zhao, and Alagan Anpalagan, "A Unified Framework for Adaptively Scheduling Hybrid Voice/Data Traffic in 3G Cellular CDMA Downlinks," WIRELESSCOM 2005, International Conference on Wireless Networks, Communications, and Mobile Computing, June 2005.

3. Jin Yuan Sun, Lian Zhao, and Alagan Anpalagan, "Soft Handoff Prioritizing Algorithm for Downlink Call Admission Control of Next-Generation Cellular CDMA Networks," PIMRC, The 16th International Symposium on Personal, Indoor and Mobile Radio Communications, September 2005.

4. Jin Yuan Sun, Lian Zhao, and Alagan Anpalagan, "Cross Layer Design and Performance Analysis of 3G Cellular CDMA Downlinks," manuscript submitted to EURASIP Journal on Wireless Communications and Networking.

Bibliography

[1] A.J.Goldsmith and S.B.Wicker, "Design challenges for energy-constrained ad hoc wireless networks," *IEEE Wireless Communications*, vol. 9, pp. 8–27, Aug. 2002.

[2] L.Alonso and R.Agusti, "Automatic rate adaptation and energy-saving mechanisms based on cross-layer information or packet-switching data networks," *IEEE Radio Communications*, vol. 9, pp. 15–20, Mar. 2004.

[3] F.Yu and V.Krishnamurthy, "Cross-layer QoS provisioning in packet wireless CDMA networks," *in Proc. IEEE Intl. Conf. Communications*, vol. 5, pp. 3354–3358, May 2005.

[4] J.Price and T.Javidi, "Cross-layer (MAC and transport) optimal rate assignment in CDMA-based wireless broadband networks," *Asilomar Conf. on Signals, Systems, and Computers*, vol. 1, pp. 1044–1048, Nov. 2004.

[5] V.Friderikos, L.Wang, and A.H.Aghvami, "TCP-aware power and rate adaptation in DS/CDMA networks," *IEE Proc. Communications*, vol. 151, no. 6, pp. 581–588, Dec. 2004.

[6] E.Hossain and V.K.Bhargava, "Cross-layer performance in cellular WCDMA/3G networks: Modeling and analysis," *in Proc. IEEE Intl. Symposium on Personal, Indoor and Mobile Radio Communications*, vol. 1, pp. 437–443, Sept. 2004.

[7] J.Yao, T.C.Wong, and Y.H.Chew, "Cross-layer design on the reverse and forward links capacities balancing in cellular CDMA systems," *in Proc. IEEE Wireless Communications and Networking Conf.*, vol. 4, pp. 2004–2009, Mar. 2004.

[8] Y.S.Chan, Y.Pei, Q.Qu, and J.W.Modestino, "On cross-layer adaptivity and optimization for multimedia CDMA mobile wireless networks," *1st Intl. Symposium on Control, Communications and Signal Processing*, pp. 579–582, 2004.

[9] C.Comaniciu and H.V.Poor, "Jointly optimal power and admission control for delay sensitive traffic in CDMA networks with LMMSE receivers," *IEEE Trans. Signal Processing*, vol. 51, no. 8, pp. 2031–2042, Aug. 2003.

[10] S.A.Ghorashi, L.Wang, F.Said, and A.H.Aghvami, "Impact of macrocell-hotspot handover on cross-layer interference in multi-layer W-CDMA networks," *Personal Mobile Communications Conference*, pp. 580–584, Apr. 2003.

[11] W.C.Jakes, *Microwave Mobile Communications*, New York, Wiley, 1993.

[12] D.Zhao, X.Shen, and J.W.Mark, "Effect of soft handoff on packet transmissions in cellular CDMA downlinks," *Proc. Intl. Symposium on Parallel Architectures, Algorithms and Networks*, pp. 42–47, May 2004.

[13] M.Soleimanipour, W.Zhuang, and G.H.Freeman, "Optimal resource management in wireless multimedia wideband CDMA systems," *IEEE Trans. Mobile Computing*, vol. 1, no. 2, pp. 143–160, Apr. 2002.

[14] S.Kim, Z.Rosberg, and J.Zander, "Combined power control and transmission rate selection in cellular networks," *in Proc. IEEE Vehicular Technology Conf.-Fall*, vol. 3, pp. 1653–1657, Sept. 1999.

[15] D.M.Novakovic and M.L.Dukic, "Evolution of the power control techniques for DS-CDMA toward 3G wireless communication systems," *IEEE Communications Surveys, fourth quarter*, pp. 2–15, 2000.

[16] J.W.Mark and S.Zhu, "Power control and rate allocation in multirate wideband CDMA systems," *in Proc. IEEE Wireless Communications and Networking Conf.*, vol. 1, pp. 168–172, Sept. 2000.

[17] S-L.Kim, Z.Rosberg, and J.Zander, "Combined power control and transmission rate selection in cellular networks," *in Proc. IEEE Vehicular Technology Conf.-Fall*, vol. 3, pp. 1653–1657, Sept. 1999.

[18] S.A.Grandhi, R.Vijayan, and D.J.Goodman, "Distributed power control in celluar radio systems," *IEEE Trans. Vehicular Tech.*, vol. 42, pp. 226–228, Apr. 1994.

[19] G.J.Foschini and Z.Milzanic, "A simple distributed autonomous power control algorithm and its convergence," *IEEE Trans. Vehicular Tech.*, vol. 42, no. 4, pp. 641–646, 1993.

[20] H.Ji and C.Y.Huang, "Non-cooperative uplink power control in cellular radio systems," *Wireless Networks (4)*, pp. 233–240, 1998.

[21] Y.J.Yang and J.F.Chang, "A strength and SIR combined adaptive power control for CDMA mobile radio channels," *Proc. Intl. Symposium on Spread Spectrum Techniques and Applications*, vol. 3, pp. 1167–1171, 1996.

[22] A.Chockalingam and L.B.Milstein, "Open loop power control performance in DS-CDMA networks with frequency selective fading and non-stationary base stations," *Wireless Networks (4)*, pp. 249–261, 1998.

[23] S.V.Hanly and D.Tse, "Power control and capacity of spread-spectrum wireless networks," *IEEE Trans. Vehicular Tech.*, vol. 35, no. 12, 1999.

[24] A.Chockalingam, "Performance of CLPC in DS-CDMA cellular systems," *IEEE Trans. Vehicular Tech.*, vol. 47, no. 3, pp. 774–789, Aug. 1998.

[25] R.D.Yates and C.Y.Huang, "Integrated power control and base station assignment," *IEEE Trans. Vehicular Tech.*, vol. 44, no. 3, pp. 638–644, Aug. 1995.

[26] F.Rashid-Farrokhi, L.Tassiulas, and K.J.R.Liu, "Joint optimal power control and beamforming in wireless networks using antenna arrays," *IEEE Trans. Communications*, vol. 46, no. 10, 1998.

[27] M.Airy and K.Rohani, "QoS and fairness for CDMA packet data," *in Proc. IEEE Vehicular Technology Conf.-Spring*, vol. 1, pp. 450–454, May 2000.

[28] S.A.Jafar and A.Goldsmith, "Adaptive multirate CDMA for uplink throughput maximization," *IEEE Trans. Wireless Communications*, vol. 2, pp. 218–228, Mar. 2003.

[29] S.J.Lee, H.W.Lee, and D.K.Sung, "Capacities of single-code and multicode DS-CDMA systems accommodating multiclass services," *IEEE Trans. Vehicular Tech.*, vol. 48, no. 2, pp. 376–384, Mar. 1999.

[30] L.Xu, X.Shen, and J.W.Mark, "Performance analysis of adaptive rate and power control for data service in DS-CDMA systems," *Proc. IEEE Globecomm Conf.*, vol. 1, pp. 627–631, Nov. 2001.

[31] L.Zhao and J.W.Mark, "Performance analysis of rate adaptation in WCDMA communication systems," *in Proc. IEEE Wireless Communications and Networking Conf.*, vol. 3, pp. 1400–1405, Mar. 2004.

[32] G.Hwang and D.Cho, "Dynamic rate control based on interference and transmission power in 3GPP WCDMA systems," *in Proc. IEEE Vehicular Technology Conf.-Fall*, vol. 6, pp. 2926–2931, Sept. 2000.

[33] S.Ariyavisitakul and L.F.Chang, "Adaptive transmission rate control scheme for ABR services in the CBR and ABR services integrated DS/CDMA systems," *in Proc. IEEE Vehicular Technology Conf.-Fall*, vol. 5, pp. 2121–2125, Sept. 2000.

[34] Y.Kwok and V.K.N.Lau, "System modeling and performance evaluation of rate allocation schemes for packet data services in wideband CDMA systems," *IEEE Trans. Computers*, vol. 52, pp. 804–814, June 2003.

[35] F.Xu and M.H.Ye, "The research on the service rate adaptation in mobile network," *Proc. Intl. Conf. on Communication Technology*, vol. 2, pp. 970–976, Apr. 2003.

[36] L.Yang and L.Hanzo, "Adaptive rate DS-CDMA systems using variable spreading factors," *IEEE Trans. Vehicular Tech.*, vol. 53, pp. 72–81, Jan. 2004.

[37] L.Wang, S.Wang, X.Sun, and Y.Liang, "Resource allocation for multimedia CDMA wireless system with soft target SIR thresholds," *Proc. Intl. Conf. on Communication Technology*, vol. 2, pp. 829–834, Apr. 2003.

[38] B.Hashem and E.Sousa, "A combined power/rate control scheme for data transmission over a DS-CDMA system," *in Proc. IEEE Vehicular Technology Conf.*, pp. 1096–1100, May 1998.

[39] V.Rodriguez and D.J.Goodman, "Power and data rate assignment for maximal weighted throughput in 3G CDMA," *in Proc. IEEE Wireless Communications and Networking Conf.*, vol. 1, pp. 525–531, Mar. 2003.

[40] F.Berggren and S.Kim, "Energy-efficient control of rate and power in DS-CDMA systems," *IEEE Trans. Wireless Communications*, vol. 3, pp. 725–733, May 2004.

[41] S.T.Chung and J.M.Cioffi, "Rate and power control in a two-user multicarrier channel with no coordination: The optimal scheme vs. suboptimal methods," *in Proc. IEEE Vehicular Technology Conf.-Fall*, vol. 3, pp. 1744–1748, Sept. 2002.

[42] C.Poulliat, L.Fijalkow, and D.Declercq, "Average performance analysis of a link adaptation strategy based on the minimum user rate maximization," *in Proc. IEEE Intl. Conf. Communications*, vol. 1, pp. 53–57, June 2004.

[43] L.Zhao and J.W.Mark, "Integrated power control and rate allocation for radio resource management in uplink wideband CDMA systems," *IEEE Intl. Symposium on a World of Wireless Mobile and Multimedia Networks*, pp. 428–436, June 2005.

[44] S.Ulukus and L.J.Greenstein, "Throughput maximization in CDMA uplinks using adaptive spreading and power control," *Proc. Intl. Symposium on Spread Spectrum Techniques and Applications*, vol. 2, pp. 565–569, Sept. 2000.

[45] D.Ayyagari and A.Ephremides, "Optimal admission control in cellular DS-CDMA systems with multimedia traffic," *IEEE Trans. Wireless Communications*, vol. 2, pp. 195–202, Jan. 2003.

[46] K.Choi and S.Kim, "Adaptive power/rate allocation for minimum mean transmission delay in CDMA networks," *in Proc. IEEE Vehicular Technology Conf.-Spring*, vol. 1, pp. 495–499, Apr. 2003.

[47] L.Song and N.B.Mandayam, "Hierarchical SIR and rate control on the forward link for CDMA data users under delay and error constraints," *IEEE J. Select. Areas Communications*, vol. 19, pp. 1871–1882, Oct. 2001.

[48] X.Duan, Z.Niu, D.Huang, and D.Lee, "A dynamic power and rate joint allocation algorithm for mobile multimedia DS-CDMA networks based on utility functions," *in Proc. IEEE Intl. Symposium on Personal, Indoor and Mobile Radio Communications*, vol. 3, pp. 1107–1111, Sept. 2002.

[49] C.Jiang and J.Lain, "Adaptive neuro-fuzzy power control and rate adaptation for multirate CDMA radio systems," *Proc. IEEE Intl. Conf. on Networking, Sensing and Control*, pp. 1307–1312, Mar. 2004.

[50] Y.Chen, Y.Lin, J.Wen, W.Chang, and J.Liao, "Combined fuzzy-based rate and selective power control in multimedia CDMA cellular systems," *in Proc. IEEE Vehicular Technology Conf.-Fall*, vol. 4, pp. 2506–2510, Oct. 2003.

[51] R.Leelahakriengkrai and R.Agrawal, "Scheduling in multimedia CDMA wireless networks," *IEEE Trans. Vehicular Tech.*, vol. 52, pp. 226–239, Jan. 2003.

[52] D.G.Jeong and W.S.Jeon, "Congestion control schemes for resverse link data transmission in multimedia CDMA systems," *IEEE Trans. Vehicular Tech.*, vol. 52, pp. 1489–1496, Nov. 2003.

[53] R.Doostnejad, E.S.Sousa, and H.Alavi, "A combined rate/power, time and sector allocation in high data rate CDMA systems based on an information-theoretic approach," *in Proc. IEEE Vehicular Technology Conf.-Fall*, vol. 3, pp. 1711–1715, Oct. 2001.

[54] A.S.Anpalagan and E.S.Sousa, "A combined rate/power/cell control scheme for delay insensitive applications in CDMA systems," *Proc. IEEE Globecomm Conf.*, vol. 1, pp. 256–260, Nov. 2000.

[55] D.Zhao, X.Shen, and J.W.Mark, "Radio resource management for cellular CDMA systems supporting heterogeneous services," *IEEE Trans. Mobile Computing*, vol. 2, no. 2, pp. 147–160, Apr. 2003.

[56] Y.Liang, F.P.S.Chin, and K.J.Ray Liu, "Joint downlink beamforming, power control, and data rate allocation for DS-CDMA mobile radio with multimedia services," *IEEE Intl. Conf. on Multimedia and Expo*, vol. 3, pp. 1455–1458, July 2000.

[57] M.Kazmi and N.Wiberg, "Power and rate assignment policies for best-effort services in WCDMA," *in Proc. IEEE Intl. Symposium on Personal, Indoor and Mobile Radio Communications*, vol. 4, pp. 1601–1605, Sept. 2002.

[58] D.Kitazawa, L.Chen, H.Kayama, and N.Umeda, "Downlink packet-scheduling considering transmission power and QoS in CDMA packet cellular systems," *IEEE Intl. Workshop on Mobile and Wireless Communications Network*, pp. 183–187, Sept. 2002.

[59] W-H.Sheen, I-K.Fu, and K.Y.Lin, "New load-based resource allocation algorithms for packet scheduling in CDMA uplink," *in Proc. IEEE Wireless Communications and Networking Conf.*, vol. 4, pp. 2268–2273, Mar. 2004.

[60] J-H.Yoon, M-J.Sheen, and S-C.Park, "Scheduling methods with transmit power constraint for CDMA packet services," *in Proc. IEEE Vehicular Technology Conf.-Spring*, vol. 2, pp. 1450–1453, Apr. 2003.

[61] J.Y.Sun, Y.F.Peng, and L.Zhao, "A novel packet scheduling scheme based on adaptive power/delay for efficient resource allocation in downlink CDMA systems," *IEEE Canadian Conf. on Electrical and Computer Engineering*, May 2005.

[62] W.E.A 3GPP 2, "1xEV-DV evaluation methodology-addendum (v6)," 2001.

[63] Y.Chen and L.G.Cuthbert, "Downlink performance of different soft handover algorithms in 3G multi-service environments," *IEEE Intl. Workshop on Mobile and Wireless Communications Network*, pp. 406–410, Sept. 2002.

[64] X.Wang, R.Ramjee, and H.Viswanathan, "Adaptive and predictive downlink resource management in next generation CDMA networks," *Proc. IEEE INFOCOM, Annual Joint Conference of the IEEE Computer and Communications Societies*, vol. 4, pp. 2754–2765, Mar. 2004.

[65] D.Wong and T.J.Lim, "Soft handoffs in CDMA mobile systems," *IEEE Personal Communications*, vol. 4, pp. 6–17, Dec. 1997.

[66] Y.Chen and L.Cuthbert, "Optimum size of soft handover zone in power-controlled UMTS downlink systems," *IEE Electronic Letters*, vol. 38, pp. 89–90, Jan. 2002.

[67] K.M.Rege, S.Nanda, C.F.Weaver, and W.C.Peng, "Analysis of fade margin for soft and hard handoffs," *in Proc. IEEE Intl. Symposium on Personal, Indoor and Mobile Radio Communications*, pp. 829–835, 1995.

[68] J.W.Chang and D.K.Sung, "Adaptive channel reservation scheme for soft handoff in DS-CDMA cellular systems," *IEEE Trans. Vehicular Tech.*, vol. 50, pp. 341–353, Mar. 2001.

[69] V.K.Garg, *IS-95 CDMA and CDMA 2000: Cellular/PCs Systems Implementation*, Prentice Hall, 1999.

[70] A.Viterbi, *CDMA Principles of Spread Spectrum Communications*, Addison-Wesley, 1995.

[71] 3GPP TS 25.331, "RRC protocol specification," 2000.

[72] 3GPP TS 25.214, "Physical layer procedures (FDD) v3.1.0," 1999.

[73] A.Viterbi, A.M.Viterbi, K.S.Gilhousen, and E.Zehavi, "Soft handoff extends CDMA cell coverage and increases reverse link capacity," *IEEE J. Select. Areas Communications*, vol. 12, pp. 1281–1288, Oct. 1994.

[74] H.Balakrishnan, S.Seshan, and R.H.Katz, "Improving reliable transport and handoff performance in cellular wireless networks," *ACM Wireless Networks*, vol. 1, no. 4, pp. 469–482, Dec. 1995.

[75] H.Balakrishnan, V.N.Padmanabhan, S.Seshan, and R.H.Katz, "A comparison of mechanisms for improving TCP performance over wireless links," *IEEE/ACM Trans. Networking*, vol. 5, no. 6, pp. 756–769, Dec. 1997.

[76] V.Jacobson, "Congestion avoidance and control," *Proc. ACM SIGCOMM*, pp. 273–288, Aug. 1988.

[77] P.Karn and C.Partridge, "Improving round-trip time estimates in reliable transport protocols," *ACM Trans. Computer Systems*, vol. 9, pp. 364–373, Nov. 1991.

[78] Y.Wu, Z.Niu, and J.Zheng, "A network-based solution for TCP in wireless systems with opportunistic scheduling," *in Proc. IEEE Intl. Symposium on Personal, Indoor and Mobile Radio Communications*, pp. 1241–1245, Sept. 2004.

[79] T.E.Klein, K.K.Leung, and H.Zheng, "Improved TCP peformance in wireless IP networks through enhanced opportunistic scheduling algorithms," *Proc. IEEE Globecomm Conf.*, pp. 2744–2748, Sept. 2004.

[80] A.Bakre and B.R.Barinath, "I-TCP: Indirect TCP for mobile hosts," *Proc. IEEE Intl. Conf. on Distributed Computing Systems*, pp. 136–143, May 1995.

[81] M.Mathis, J.Mahdavi, S.Floyd, and A.Romanow, "Selective acknowledgment options," *RFC-2018*, 1996.

[82] S.Floyd, "TCP and explicit congestion notification," *ACM Computer Communication Review*, vol. 24, pp. 10–23, Oct. 1994.

[83] S.Lin and D.J.Costello, *Error Control Coding: Fundamentals and Applications*, Englewood Cliffs: NJ: Prentice-Hall, 1983.

[84] "The network simulator - ns-2," *http://www.isi.edu/nsnam/ns/.*